A practical guide to orthology resources

Paul de Boissier, Bianca Habermann

A practical guide to orthology resources

Paul de Boissier and Bianca H. Habermann

Aix-Marseille University, CRNS, IBDM UMR 7288, Marseille, France

Corresponding author :
Bianca H. Habermann
Computational Biology Group
IBDM – Institut de Biologie du Développement de Marseille (UMR 7288)
CNRS & Aix-Marseille Université
Case 907 – Parc Scientifique de Luminy
163 Avenue de Luminy
13009 Marseille
France

e-mail: Bianca.HABERMANN@univ-amu.fr
phone: +33 (0)4 91 26 92 36

Abstract

Many resources exist which collect information on orthologous genes and provide this information to the research community. However, the algorithms used to collect orthologs, the type of data available for analysis or download, the range of organisms included, as well as the user-friendliness vary greatly between different orthology databases. In this review, we present a practical guide to the best-known orthology resources: we here briefly discuss their algorithmic details, review their taxonomic coverage and illustrate their user-friendliness. Moreover, we evaluate their capability to detect remotely conserved orthologs and to resolve inparalog relationships in gene families. Moreover, we test them for potential false-positive classification by using a multi-domain protein family with a complex evolutionary history. Finally, we assess the availability and ease of usage of orthology search engines offered by orthology database providers for local usage.

Introduction

With Next Generation Sequencing (NGS) methods, the number of completely sequenced genomes - and thus the availability of complete proteomes - has increased tremendously (Figure 1). One essential step after genome sequencing is to annotate its gene products and to predict the putative functions of an organism's proteins. The most common method for functional annotation is to infer a protein’s function from its related sequences, namely its orthologs from other, already annotated species. The fundamental basis of the concept for transferring functional information across orthologs is the ‘ortholog conjecture’ or the standard model of phylogenomics (Koonin 2005; Altenhoff et al. 2012) . This theory states that orthologs retain the ancestral function, while paralogs tend to rapidly evolve novel functions (Altenhoff et al. 2012) . As many organisms will not be studied experimentally, the functional annotation of their genomes relies exclusively on transferring functional knowledge on proteins from other, experimentally studied organisms.

The formal definition of homologous proteins is that they share a common ancestor, and thus, are homologous on sequence level. Homologs can be divided in different categories depending on their ancestry (Koonin 2005) : Orthologs, which will be discussed here, result from an event of speciation. Orthologs are typically used to infer gene functions for newly sequenced species. Paralogs are homologous proteins resulting from a gene duplication event. These can be further divided into inparalogs, which result from gene duplication after speciation; and outparalogs, which result from a duplication event before speciation (Figure 2). Finally, xenologs result from horizontal gene transfer. Gene duplication and gene losses, together with horizontal gene transfer make the distinction of orthologs often difficult, as it is sometimes hard to distinguish, whether a predicted ortholog has arisen from speciation, or from a combination of gene duplication and gene loss events.

There are in principle two types of approaches for identifying orthologs: phylogeny-based methods and methods based on the Reciprocal Best Hit (RBH) theory (Wolf and Koonin 2012) . Performing a phylogenetic analysis requires to collect family members, align them, calculate a phylogenetic tree and reconcile the tree for gene gains and losses. Phylogeny-based orthology inference methods tend to be more accurate, as they require a certain amount of manual curation, such as optimizing multiple-sequence alignment, and offer a wider choice of parameters, e.g. for tree reconstruction. However, this makes it also harder to compare different phylogeny-based orthology resources (Kriventseva et al. 2008) . Furthermore, phylogeny-based methods tend to be computationally expensive. RBH-based methods (which can also be referred to as best reciprocal hit (BRH) or best-best hit (BBH) or genome-specific best hit (BeT)) rely on sequence similarity searches and consider two proteins orthologous if they are each other's best hit in their respective proteomes. They were first introduced with the cluster of orthologous groups (COG) database (Tatusov et al. 2003). RBH-based methods are easy to implement computationally and can be scaled up to treat hundreds and thousands of genomes. Thus, RBH-based methods made it possible to automatize orthology assignment and thus they are at the base of many orthology search engines published to date.

Several tools and databases were created to group and unify genes and proteins based on their evolutionary relationship. These orthology resources are very useful in guiding biologists of different disciplines through the evolutionary history of their proteins of interest. As they use different approaches to collect orthologous proteins, contain different sets of organisms and offer different analysis tools, the information they provide and their user-friendliness differs substantially.

In this chapter, we will focus on orthology resources and aim at helping the reader to find a suitable database for identifying orthologous genes. We will discuss their user-friendliness, their completeness and whether they can resolve problems caused by inparalogs and remote orthologs.

One obstacle in the quest for orthologs is remote orthology. Remote orthologs typically share below 20% sequence identity at protein level; this zone is referred to as the twilight zone of sequence similarity (Walter 1989). Thus they are difficult to detect with traditional search methods such as BLAST. To discover remote orthologs, more sensitive methods such as profile-based methods have to be used (Steinegger et al. 2019). It is therefore not surprising that remote orthologs are not detected by many orthology search engines. Yet, some search engines do manage to include more remotely conserved orthologs when identifying gene families. In order to probe orthology resources and their underlying algorithms for their ability to detect also remote orthologs, we have selected the cytochrome C oxidase assembly protein COX20 from *Homo sapiens* (Table 1). Human COX20 (aka FAM36A) was found as a remote ortholog of the protein COX20 of the budding yeast *Saccharomyces cerevisiae* (Szklarczyk et al. 2013). Human COX20 is only half the size of its yeast ortholog (118 vs 205 amino acids (aa), respectively). Submitted to Needle of the EMBOSS software from the EBI (Rice et al. 2000) , with a gap open penalty of 10 and an extend penalty of 0.5 and the BLOSUM62 matrix, these two proteins share only 27% of sequence similarity and 14% sequence identity, which makes them remote orthologs (Figure 3 a). We can therefore use COX20 to estimate the completeness of different databases with respect to proteins with low sequence conservation.

We also wanted to assess orthology resources for their ability to resolve inparalog relationships. Thus, we selected pyruvate carboxylase protein (PC) of *H. sapiens* as our second test case. PC is known to have two inparalogs in *S. cerevisiae* (Pronk et al. 1996), PYC1 and PYC2 (Table 1). When submitted to Needle using the same parameters as were used for COX20, human PC has 68.2% sequence similarity to PYC1 from *S. cerevisiae*; and 68.4% sequence similarity to yeast PYC2, respectively. PYC1 and PYC2 are more than 97% similar to each other, which makes them inparalogs, being more similar to each other than to their ortholog(s) in another species (Figure 3 b).

Finally, we wanted to investigate putative false-positive assignments. Best candidates for potential false-positive classifications are multi-domain proteins in outparalog relationships. We chose the tailless protein family, which contains a nuclear hormone receptor (NR) domain together with a Zinc finger domain. Tailless from *Drosophila melanogaster* has three close paralogs, tailless (tll), dissatisfaction (dsf) and hormone receptor 51 (Hr51). Two human proteins, NR2E1 and NR2E3 are equally member of this NR sub-family (Figure 3 c). While it is difficult to unequivocally assign orthology in multi-branching families, phylogenetic analysis, in agreement with many orthology resources, assigns NR2E1 as orthologous to tll and NR2E3 as orthologous to Hr51. Needle from EMBOSS reports 51.5% sequence similarity between tll and NR2E1, and only 33.2 % sequence similarity between dsf and NR2E1, mostly owing to the fact that the dsf protein contains a long insertion in the center of its sequence and is thus 239 amino acids longer than tll.

OrthoDB

The first database we are discussing is OrthoDB (Kriventseva et al. 2008; Kriventseva et al. 2019) . It is referred to as a “catalog of orthologs” and computes orthologs on various levels of the taxonomic hierarchy. OrthoDB relies on the RBH method. It first finds best hits between species using the very fast and sensitive MMseqs2 algorithm (Steinegger and Söding 2017) . Clusters of orthologs are then build progressively, with specific e-value cut-offs for triangular RBHs and bidirectional RBHs. Clusters are then further expanded to include inparalogs that are identified as more similar to each other within species than to any protein in another species. OrthoDB has since its introduction embraced the fact that orthologous groups are hierarchical. The procedure to identify orthologs is thus applied at each major radiation of the species taxonomy. As a result, it produces more finely resolved groups of closely related orthologs. Functional annotations are added to each group by summarizing the respective annotations from UniProt, NCBI Gene, InterPro and Gene Ontology. In January 2020, it contained data for 1271 eukaryotes, 6013 prokaryotes (5609 bacteria and 404 archaea) and 6488 viruses for a total of 37 million genes.

To search OrthoDB, the user can perform a simple text search, use identifiers from various databases or a protein sequence. The sequence search is limited to 1000 amino acids, which makes it impossible to search with large protein sequences, such as Titin (~30000 aa). The advanced search option allows adding specific species to the search, which are presented in a tree-like interface. In the simple text search, the user can specify, if the gene has to be present in all species, in more than 90% or 80%; and if it has to be present in a single copy in all, or more than 90% or 80% of species.

When performing a simple text search with the term ‘Cox20’, OrthoDB returns 246 groups, corresponding to search hits at different taxonomic levels, from the level ‘Eukaryota’ down to sub- and even infraorder levels. The user can thus easily mine orthology relationships at wide taxonomic range. The identifier from the NCBI database must be the GeneID (here 116228). The results are more precise as it returns only the COX20 group. The same applies to the sequence-based search. At the Eukaryota level, the COX20 group contains 922 orthologs in 875 species. The Group hierarchy is shown in an interactive plot right at the top of the web-page (Figure 4 a). The annotation of the protein family includes a functional description, Gene Ontology (GO) terms, and the evolutionary descriptions, including number of copies per organisms, evolutionary rate, and gene architecture (Figure 4 b). Orthologs by organisms are listed in the third part of the page, with a link to each protein entry at UniProt and InterPro (Figure 4 c). At the bottom of the page, the sibling groups are listed with % overlap and InterPro domains (not shown).

Searching for pyruvate carboxylase (PC) orthologs was done using the GeneID. The search could not be performed using the sequence, as the protein has more than 1000 aa. A text-based search with the gene name ‘PC’ is less accurate, as many groups contain these two consecutive characters. Search results for the PC GeneID (5091) revealed that OrthoDB is able to resolve inparalog relationships, as the database returned PC for *H. sapiens* and

PYC1 and PYC2 for *S. cerevisiae*. Only the naming of the group in OrthoDB is confusing, as it is referred to as *Biotin carboxylase, C-terminal* in the Eukaryota group, and to *Pyruvate Carboxylase* starting from Metazoa. In total, this group contains 3888 genes found in 1188 species.

We searched for tailless with its GeneID from NCBI (43656), which returned in Eukaryota the Nuclear Hormone Receptor, ligand binding domain group (870262at2759) with 1305 genes in 444 organisms, including the three *Drosophila* (tll, dsf and Hr51) and 2 human proteins (NR2E1, NR2E3). More detailed information on the relationship between these 5 proteins is not returned.

In conclusion, we find that this tool provides extensive information and accurate orthology assignments. It is fast and user-friendly. It succeeded in finding the remote ortholog of COX20 and included the inparalogs from the pyruvate carboxylase family. OrthoDB does not provide detailed information on the phylogenetic relationship of the 5 proteins of the tailless family, however correctly and exclusively identified them as being part of the same group. Search options are manifold, though the easiest and most precise results are returned when searching with either with the NCBI GeneID, or the sequence. OrthoDB is available at: https://www.orthodb.org/

HomoloGene

HomoloGene (NCBI Resource Coordinators 2016; NCBI Resource Coordinators 2018) is a tool developed by the NCBI to detect paralogs as well as orthologs. It contains 21 completely sequenced Eukaryotic genomes and profits from the entire information content of the NCBI databases, including in-depth information provided for annotated genes at NCBI (synonyms, gene description, genomic location, isoforms, Gene Ontology (GO) information, interaction partners, or literature). HomoloGene uses sequence similarity based on BLASTp (Altschul et al. 1997) comparisons to match sequences into groups using a species tree. More closely related sequences are matched first, followed by more distantly related ones. The algorithm for sequence matching is heuristic and performs bipartite matching, an algorithm derived from graph theory (Bondy and Murty 1976). The matching procedure employed by HomoloGene optimizes the global, rather than the local score of the bipartite graph. For each match, a statistical significance is calculated. Protein alignments are mapped back to their respective DNA sequences to obtain Ka/Ks ratios: the ratio of the number of substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) over a given time-frame. These Ka/Ks ratios are used to filter out sequences that have potentially been incorrectly grouped. Inparalogs are determined by identifying sequences that are more closely related to a sequence within one organism than to a sequence in another organism.[1]

For searching HomoloGene, all basic and advanced search options are available, such as searching different fields, AND/OR options, drop-down lists, etc. We used the basic search for COX20, which returned two HomoloGene groups: an ‘unnamed protein’ group in *Saccharomycetaceae*, which corresponds to COX20 in *S. cerevisiae* and closely related fungi; and the ‘Cox2 chaperone homolog (*S. cerevisiae*)’ group conserved in *Euteleostomi*, corresponding to COX20 conserved from human to zebrafish (Figure 5 a). Protein identifiers, species and gene names are shown. Additionally, proteins are represented graphically with

[1] It should be noted at this point that information on the HomoloGene algorithm could not be found even in the pages of the NCBI help desk. Information on the build procedure of HomoloGene is therefore taken from Wikipedia (https://en.wikipedia.org/wiki/HomoloGene).

their conserved domains. Identifiers are linked to the Gene entry and the Refseq protein entry, respectively; the graphical protein representation is linked to the conserved domain database (CDD, (Lu et al. 2020)) entry of the respective domain(s). Users can view and download the multiple sequence alignment of all family members, which can be useful for further analysis; BLASTp based pairwise alignments of orthologous proteins (Figure 5 b and c) can also be downloaded for further analysis or processing. Individual alignment scores for all pairwise comparisons are accessible from the link “Show Pairwise Alignment Scores”. Finally, there is an exhaustive list of articles linked to proteins from the orthologous group.

The search for inparalogs revealed that HomoloGene groups all paralogs in the same orthology group. When searching for pyruvate carboxylase, the two different proteins of *S. cerevisiae* were part of the same group (Figure 5 d). PC has several domains, which are shown in different colors.

HomoloGene does not provide information on the entire tailless family, but rather classifies orthologous pairs. Interestingly, and opposing to phylogenetic analysis and other orthology resources, HomoloGene assigns dsf as orthologous to NR2E1. Tll is classified as conserved in Diptera only and Hr51 is assigned as the ortholog of NR2E3. A better Ka/Ks ratio could be causative for defining dsf as the ortholog of NR2E1, as there is locally a higher number of identical amino acids in the ZnF and NHR domains (Figure 5 e).

In conclusion, the main advantage of HomoloGene is its integration in the NCBI database resources and the high confidence on orthologs inherent in its build procedure. Furthermore, inparalogs are identified and indicated as both being part of one HomoloGene group. The ambiguous tailless family is resolved differently than in other orthology databases: dsf, instead of tll is considered the ortholog of NR2E1. Tll itself is classified as arthropod-specific and Hr51 is considered orthologous to NR2E3. Moreover, HomoloGene does not provide an overview of the entire tailless-family, but rather separates tll, dsf and Hr51 into three different families. Disadvantages include the low number of organisms (21 species versus 13722 e.g. in OrthoDB) and the disability to identify remote orthologs. HomoloGene has not been updated since 2014, according to information provided at ‘HomoloGene statistics’. HomoloGene is available at: https://www.ncbi.nlm.nih.gov/homologene/

TreeFam

TreeFam (Ruan et al. 2008; Schreiber et al. 2014) is one of the bioinformatic tools developed at the EMBL-EBI. The last release (v9, from March 2013) contains 15736 gene families from 109 species. TreeFam v9 has adopted the Ensembl Compara pipeline to assemble ortholog families, which performs all-agains-all BLASTp searches and subsequent clustering of ortholog families. The clustering procedure in TreeFam however uses a Hidden Markov Model (HMM-) based approach to cluster TreeFam families, which allows the database to have more stable ortholog families with new releases (Schreiber et al. 2014), as new sequences can be added to existing HMM-families. Multiple sequence alignments are created with either MAFFT (Katoh and Standley 2013) or MCoffee and refined by removing non-conserved positions. TreeBest is used to construct gene trees (http://treesoft.sourceforge.net/treebest.shtml). Several trees (based on amino acid as well as back-translated nucleic acid alignments) are constructed and a consensus tree is calculated using a species tree as a reference. Gene losses and duplications are calculated using the Duplication/Loss Inference algorithm (Li et al. 2006) and by reconciling the tree with the NCBI taxonomy tree (Federhen 2012) .

TreeFam can be browsed with a gene name, or searched with a protein sequence. By clicking on the TreeFam family, first a summary page is invoked, which gives the user information on the general conservation of the query in a species tree. The Gene Tree tab (Figure 6 a, found on the right-hand side of the page) displays the gene tree, as a model, which can be expanded (by clicking on 'Full'); the tree can be annotated by adding information on branch length, bootstrap values, labels for taxonomy, etc. The protein nodes are linked to the gene entries of the Ensembl database (Cunningham et al. 2019). Next to the tree are graphical representations of the proteins with identified conserved domains. These link to the respective entry of the conserved domain in the Pfam database (El-Gebali et al. 2019) . Wikipedia links to the Wikipedia entry of the gene; the sequences in the tree can be assessed at the 'Sequences' link; finally, the alignment in fasta format, the HMM, as well as the tree in Newick format can be downloaded from the 'Download' link. Summary of the gene family is displayed at the top of the entry (listing number of species, sequences, alignment length and % overall identity). When searching with a protein sequence, the sequence will be grouped to its presumable family by similarity and added to the family tree. The user can choose between two phylogenetic methods: Parsimony which is less accurate but faster and Maximum Likelihood which is slower but more accurate. The sequence is added at the correct position in the tree, however as a separate - duplicated - entry. For instance, when searching with the COX20 protein sequence from human, there will be two identical human nodes in the tree.

COX20 of *S. cerevisiae* is not found in the pre-built tree, though when searching with the budding yeast COX20 protein sequence, the correct family is identified and the sequence is added to the tree.

When searching for the PC family, the two inparalogs in *S. cerevisiae* are correctly placed in the family tree of pyruvate carboxylase. Moreover, the Gene Tree tab allows the user to see, in a small synthetic view with model organisms, where the duplication and speciation events occurred. The red triangle and the green point respectively represent the duplication and the speciation events (Figure 6 b).

The tailless family contains next to tll, dsf and Hr51 also the protein Hr83 from *Drosophila* (Figure 6 c). This protein is not classified as part of the tll family in other databases.

To summarize, this database is quite comprehensive, providing visual display of the family tree, allowing download of underlying alignments and trees and providing functional annotation from Wikipedia. The tree is interactive and can be labeled. Genes are linked to the Ensembl resource, providing rich information on individual genes/proteins. A novel sequence can be added to the family tree. It is able to distinguish inparalogs, in fact indicating speciation and duplication events in the tree. Moreover, tree-based methods are thought to be more powerful in inferring orthology than simple RBH-based approaches (Koonin 2005; Brown and Sjölander 2006) . It however did not include *S. cerevisiae* COX20 in the pre-calculated COX20 tree. TreeFam also fails to correctly distinguish orthology relationships of the tailless family, as it assigns false-positive ZnF and NHR-domain containing proteins to this family. Another disadvantage is the lack of recent updates of the database. TreeFam is available at: http://www.treefam.org/

HCOP

HCOP (HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions (Eyre et al. 2007)) is a database of 19 species, including *Homo sapiens* and *Saccharomyces cerevisiae,* which allows identification of pairs of predicted orthologs. It combines orthology data from 14 different databases of orthologs, including OrthoDB, OrthoMCL, HomoloGene, OMA, TreeFam, Panther, Inparanoid, EggNOG and others. Thus, it integrates orthology data derived from many different search strategies (Figure 7 a). Orthologous pairs from these sources are consolidated into a non-redundant list of orthologs and HCOP provides the associated list of databases that support each assignment. HCOP is human-centric as orthologs of a human protein can be found in other species but searching for a protein of another species only returns orthologs in *H. sapiens*.

The search can be done by identifiers of Ensembl, NCBI or HGNC, approved gene symbols or a file containing a list of identifiers. Wildcards can be used : "_" to substitute a single character and "*" or "%" for zero, one or several characters. We used the approved symbol COX20. The output is a list of orthologs in the different species which can be saved as a text file. The results give the chromosomal location, and specific identifiers link the ortholog to the database it is found in. The support from the different orthology resources is furthermore shown for each identified ortholog. Budding yeast COX20 is found as an ortholog and supported by the PANTHER database (as indicated by the PANTHER symbol in the search results). Searching with *S. cerevisiae* COX20 only retrieves the human ortholog (Figure 7 b). Searching PC using its official gene name returns all orthologs and inparalogs, including PYC1 and PYC2 from *S. cerevisiae*.

When searching for tll in HCOP, the two human proteins NR2E1 and NR2E3 are found, suggesting that HCOP groups the entire tailless family. When searching for orthologs of human NR2E1, tll, dsf and Hr51 are identified in fruit fly.

In summary, the HCOP database is a very useful resource as it offers cross-references between different orthology databases. As it relies on data generated by many different search algorithms, it is able to find remote orthologs and includes inparalogs in orthology groups. HCOP limits the tailless family in human and *Drosophila* to the core members. Furthermore, all data can be downloaded in tabular format for local usage. The main disadvantage is that there are only few species included; moreover, it is human-centric and only returns all orthologs when searching with the human sequence. HCOP is available at : https://www.genenames.org/tools/hcop/

OMA

OMA (Orthologous Matrix) was developed at and is hosted by the Swiss Institute of Bioinformatics (SIB) (Altenhoff et al. 2015; Altenhoff et al. 2018). To infer orthologues, it first computes all-against-all Smith-Waterman alignments, saving only candidate pairs with sufficient score and overlap. In the next step, evolutionary distances are used to identify closest homologs, thus defining orthologs based on the reciprocal best hit hypothesis, however considering potential gene losses. Identified orthologs are finally clustered into OMA groups (i.e. most closely related genes between each two species and thus contain only orthologs), which tend to be very specific, as well as hierarchical orthologous groups (HOGs, which are hierarchical groups of all genes that descended from a single common ancestor and thus contain (in)paralogs). A detailed primer of the OMA database and search algorithm is given in a recent review by Zahn-Zabat, et al. (Zahn-Zabal et al. 2020) . OMA is actively maintained (as of 2020) and contains 2288 species (1674 Bacteria, 152 Archaea and 462 Eukaryotes (fungi, animals and plants)). Model organisms are updated at each release and other genomes are updated at each important re-annotation or added based on user requests. OMA provides domain annotations and synteny data for each gene; moreover, Gene Ontology (GO) terms are inferred for each cluster.

OMA can be searched with either one recognized identifier, an amino-acid sequence, an OMA group or by performing a simple text search. Searching with a sequence either performs an 'exact' search, returning only hits that match the input sequence exactly; or an 'approximate' search, where a few mismatches are allowed. The approximate search is not comparable to BLAST, but rather used for retrieval of near-identical entries in the database. An approximate search with the protein sequence of human Cox20 returned 127 entries of high homology, representing COX20 1:1 orthologs. When searching with the term COX20 in a free-text search, a list of all COX20 proteins is returned. The user can choose to either see the protein record; or go directly to identified orthologs. *H. sapiens* COX20 has 139 one-to-one orthologs. A tabs-menu lets the users switch from tabular listing of one-to-one (1:1) and one-to-many (m:1) orthologs (Figure 8 a), to information on the protein, to local synteny (Figure 8 b), to OMA group (downloadable in fasta format) and HOGs (Figure 8 c). By invoking the 'OPTIONS' drop-down menu, boxes next to genes in the tree can be colored according to the gene length or the % CG content. Boxes are also interactive, linking to the HOG table of the gene, as well as the sequence (Figure 8 d). An alignment can be created, visualized, filtered and downloaded for each OMA group (Figure 8 e). Furthermore, a fingerprint is created for each group, representing the most conserved region of its members (Figure 8 e, top). For each species, close OMA groups (Figure 8 f) and gene ontologies (Figure 8 g) are listed in an easily readable tab format.

OMA was not able to find the remote ortholog from *S. cerevisiae* and the COX20 OMA and HOGs group were confined to metazoans. PC, though present in the database, is not grouped in any OMA or HOG cluster. The two inparalogs PYC1 and PYC2 have nearly 800 1:1 orthologs; they are part of an OMA group of 67 members from Bacteria and Eukaryotes. The two inparalogs are listed in this group as "close paralogs".

The tailless family is split in separate orthologous groups in OMA. Tll itself is grouped with human NR2E1, Hr51 is found in the same OMA group as human NR2E3. Dsf is classified as an arthropod-specific protein.

In summary, OMA can be considered a database with a rich visual interface, providing plenty of information and harboring many species. The output is well integrated and visualized. Information on local synteny of a gene of interest should be mentioned, as it is not found in many other orthology databases and can lead to the discovery of gene clusters implied in one mechanism. This is specifically relevant for bacterial proteins. However, OMA was neither able to identify the COX20 orthologs from human and *S. cerevisiae*; nor was it correctly classifying PC with budding yeast PYC1 and PYC2. Nonetheless, these two were correctly identified as close paralogs. Finally, the tailless family was divided in separate, orthologous groups. OMA is available at : https://omabrowser.org/oma/home/

OrthoMCL DB

OrthoMCL DB (Li et al. 2003; Chen et al. 2006; Fischer et al. 2011) is a part of the EuPathDB project and relies on the OrthoMCL clustering algorithm to identify orthologs. Orthologs are identified using WU-BLASTP (Altschul and Gish 1996) and the RBH strategy, using an e-value better than 1e-5 as a cut-off for identifying orthologs. Retained orthologs – as well as inparalogs – are linked in a network of orthologs. Edges connecting nodes (orthologs) are weighted using BLAST similarity scores. A graph-based cluster algorithm, the Markov Cluster algorithm (MCL) (Enright et al. 2002) is used to create groups of orthologs. In brief, MCL performs random walks on graphs using Markov matrices to calculate transition probabilities from one node to the other. This graph-based clustering algorithm is less computationally expensive than tree-based methods for clustering orthologs. OrthoMCL DB contains 150 species (36 Bacteria, 16 Archaea and 112 Eukaryotes) and was last updated in July 2015.

The search can be done by OrthoMCL DB IDs, free text search, a phyletic pattern, by function, by groups or by sequence. Searching for the synonym of human COX20, FAM36A, 21 orthologs were found mostly in Metazoans. This information is displayed in a simple, colored tabular format on the results page, where abbreviations of species name are associated with a 0 (not found in this species) or a 1 (found in this species) (Figure 9 a). The ortholog in *S. cerevisiae* is not found. An interesting feature of OrthoMCL DB is the display of orthologous groups as a network (Figure 9 b). This allows interactive visualization of orthologs and how they are related to each other. More or less stringent cut-offs can be chosen to reconstruct the orthology graph. When searching with the text term COX20, 16 orthologs are found, mostly in fungi (Eukaryotes). The ortholog in *H. sapiens* is not found for the fungal groups. Information about taxon, identifiers and domain architecture is provided for each gene and species.

Pyruvate carboxylase is classified with other carboxylases. Three proteins are found for *S. cerevisiae* and *H. sapiens*, respectively*.* These include in human PC, a Propionyl-CoA carboxylase and a Methylcrotonoyl-CoA carboxylase; in budding yeast, PYC1, PYC2 and an Urea amidolyase is included. These different proteins can be subdivided in different orthologous groups. In case of PC, the Enzyme Commission (EC) number can be used to identify the correct enzyme. The EC number of pyruvate carboxylase is 6.4.1.1, meaning it is a part of the ligases (6), forming carbon-carbon bonds (6.4) so ligases that form carbon-carbon bonds (6.4.1), thus, it is a pyruvate carboxylase (6.4.1.1).

The tailless family is split in separate groups, one encompassing tll and NR2E1 and the second one containing Hr51 and NR2E3. We could not find an OrthoMCL group for dsf with any of the valid identifiers, gene names or synonyms.

In conclusion, while OrthoMCL provides a fast, graph-based algorithm to cluster orthologs derived from RBHs, the algorithm is not able to identify distant orthologs; nor is it useful when searching for proteins that are part of large superfamilies. Inparalogs are classified in OrthoMCL but, with the implemented clustering algorithm, different orthologs can be clustered in superfamilies, which can be disturbing for a user who is looking for a specific protein. As also observed in other orthology resources, the tailless family is divided in two separate groups. OrthoMCL is available at : https://orthomcl.org/orthomcl/

P-POD

P-POD (Princeton Protein Orthology Database (Heinicke et al. 2007; Livstone et al. 2011)) uses all-against-all BLASTp searches followed by two alternatives types of clustering methods to group orthologous proteins: predicted orthologs based on the OrthoMCL clustering algorithm; and larger protein families that are clustered based on a Jaccard index (Jaccard 1912) inferred from shared sequence similarity, putting an orthologous group in its larger evolutionary context. Multiple sequence alignments and phylogenetic trees are created for each group using MAFFT (Katoh and Standley 2013) and PhyML (Guindon et al. 2010), and gene loss and duplication events are resolved using Notung (Chen et al. 2000) . Information from species-specific databases are collected for genes, organisms and diseases. It offers information for 12 model organisms including, *H. sapiens*, *S. cerevisiae or Arabidopsis thaliana*.

Only a text search is possible, using either gene name, IDs or an OMIM (Online Mendelian Inheritance of Man) ID. A search with COX20 returns the yeast protein, which is classified as a sequence orphan. When searching for the synonym of human COX20, FAM36A, P-POD found orthologs in different species, including *Drosophila melanogaster*, but not in *S. cerevisiae*. Looking for Pyruvate carboxylase in *H. sapiens* allows to find the two inparalogs in *S. cerevisiae*. According to the method used, different results are found. Using Jaccard clustering, 7 proteins are found in *H. sapiens* and 4 in *S. cerevisiae*, which include other members of this superfamily (Figure 10 a). A tree representation can be accessed, showing the evolution of the gene and the possible events of duplication as calculated by the Notung package (Chen et al. 2000) (Figure 10 b).

Results for the tailless family show inconsistencies with other resources. There are two distinct PPOD groups for tll (with human NR2E1) and Hr51 (with human NR2E3). The Multi/InParanoid- and naïve Ensemble- Hr51 groups both contain additionally the *Drosophila* Hr83 protein. The Multi/InParanoid group of tll contains in addition the *Drosophila* dsf protein, as well as 2 human proteins that are not considered part of this family in other resources, namely NR2F1 and NR2F2. The naïve Ensemble group of tll contains Drosophila proteins tll, dsf and seven-up (svp), as well as the human proteins NR2E1, NR2F1 and NR2F2.

In conclusion, P-POD is less comprehensive than other orthology databases. It only is available for a limited number of model organisms. Due to the stringency of its algorithm, it does not find remote orthologs. The tll family in PPOD includes moreover a number of false-positive proteins. However, as it includes Notung as one step in its analysis pipeline, it is well suited to resolve gene losses and duplications and thus to correctly identify inparalogs. P-POD is available at: http://ppod.princeton.edu/

InParanoid

InParanoid (O'Brien et al. 2005; Sonnhammer and Östlund 2015) uses reciprocal BLASTp searches to identify orthologs via the RBH method. In Version 8, there are 273 proteomes in the database (246 Eukaryotes, 20 Bacteria and 7 Archaea), extracted from Ensembl and UniProt. Inparalogs are separated in the output and outparalogs are excluded. The user can set the score for excluding inparalogs when invoking an InParanoid search.

InParanoid offers several different search options, including text-search, identifier search, or a sequence-based search. When searching for COX20 in human, all pairwise groups of orthologs are returned, which makes navigation of results difficult (Figure 11 a). The budding yeast COX20 ortholog is not found. Likewise, using the human COX20 sequence for the search, *S. cerevisiae* COX20 is not found. Searching for FAM36A even in a text search returns orthologs of human COX20, however not the human protein itself, which indicates that alternative identifiers are not supported. When searching for COX20 in *S. cerevisiae*, only its orthologs in other fungi are found.

InParanoid is designed for identifying inparalogs, so it is not surprising to find members of the pyruvate carboxylase correctly (Figure 11 b). Using either PC or PYC is non-practical as text-search, as the database cannot disambiguate the name of this gene. Searching in the gene search for the gene name PC returns mostly clusters of pyruvate carboxylase, yet includes also polycomb protein c (Pc) from *Drosophila melanogaster*.

Like HomoloGene, InParanoid groups *Drosophila* dsf with human NR2E1. Tll is in a separate InParanoid cluster, not containing a human ortholog. Hr51 is clustered with NR2E3.

In conclusion, InParanoid offers a moderate range of organisms and is limited to well-conserved orthologs that can be found by BLASTp. Inparalogs are resolved correctly and the user can choose a score to include or exclude inparalogs. The tailless family is split in three clusters, whereby dsf is considered orthologous to NR2E1, not tll. The output of InParanoid is pragmatic and simple, however non-practical, as each cluster the query is found in, is shown and no family cluster is created for proteins belonging to the same orthologous group. InParanoid is available at : http://inparanoid.sbc.su.se/cgi-bin/index.cgi

KEGG orthology Database

The KEGG orthology Database (Kanehisa et al. 2014; Kanehisa et al. 2016a; Kanehisa et al. 2017) is a part of the Kyoto Encyclopedia of Genes and Genomes (KEGG). It contains at least 4000 genomes. Orthology data are collected using KOALA (KEGG Orthology And Links Annotation (Kanehisa et al. 2010)). This tool evaluates similarity scores, best-hit relationships, domains and taxonomy to assign genes and proteins to a group of orthologs. The data are also verified manually using experimental evidence and literature. The orthologs are classified in groups by a specific KOALA number (K number). The orthology groups are fully integrated with the rest of the KEGG resources, for example by linking to KEGG Genes. KOALA groups are linked to BRITE hierarchies and KEGG pathway maps. The KOALA group can also be displayed as a simple hierarchy, following the NCBI taxonomic classification.

The remote ortholog of *S. cerevisiae* is found for human COX20. These proteins belong to the large group K18184 with over 300 members. The orthology information is displayed in KEGG style, with Brite functional annotation of the orthologous group. COX20 for instance belongs to the group 09150 associated with Organismal Systems and the group 04714 associated with Thermogenesis. The genes that are part of group K18184 are listed right below the Brite hierarchy and links to the KEGG organisms within the NCBI taxonomy, the KOALA list of genes, as well as the UniProt list of genes. The last part of the box contains information on the literature and the sequence entry itself. Next to the main box, all links within KEGG are shown (Figure 11 c).

The name ambiguity of pyruvate carboxylase (PC) again gave a long list of results, whereas looking for PYC gave only five results. Thus, it is best to search for pyruvate carboxylase or to use an identifier accepted by KEGG. PC belongs to the group K01958 and the two *S. cerevisiae* inparalogs are correctly identified. The EC commission (IUBMB) has assigned the EC number 6.4.1.1.

KEGG separates the tailless family in three KO groups. Tll is clustered with human NR2E1, Hr51 with human NR2E3. Finally, dsf does not have a human ortholog.

In conclusion, KEGG Orthology is a highly interlinked database, which can take advantage of all information available by KEGG including information on genes, pathways, ontologies, disease and literature. The database is not very visual, except for the available pathway maps, which makes browsing the results somewhat difficult. Its annotation strategy however correctly identified the remote ortholog COX20 from *S. cerevisiae* and is also able to resolve inparalog relationships. The tailless family is split into three groups, with the orthology assignments following the predominant consensus of other orthology resources. KEGG orthology Database is available at: https://www.genome.jp/kegg/ko.html

EggNOG

EggNog (Evolutionary Genealogy of Genes : Non-supervised Orthologous Groups) was developed by the Computational Biology team at the EMBL in Heidelberg (Jensen et al. 2008; Huerta-Cepas et al. 2019). We describe here the latest available pipeline, which consists of many steps. Reciprocal hits are derived from all-against-all Smith-Waterman alignments provided by the SIMAP project (Arnold et al. 2014). In the next step triangular clustering – searching for reciprocal best hits using sets of three species – is performed to identify (OG). In this pipeline, inparalogs are identified and treated as one sequence to ensure that they finally belong to the same cluster. Each OG is annotated using a functional annotation pipeline that consolidates the annotations of the identified species within an OG. Similar to OrthoDB, different taxonomic levels are used to compute OGs independently. This ensures a more accurate functional annotation. These nested OGs are tested and corrected for consistencies. A phylogenetic tree is calculated for each OG using the python-based ETE pipeline (Huerta-Cepas et al. 2016) . In brief, multiple sequence alignments are created using Clustal Omega (Sievers et al. 2011) and soft trimmed to remove columns with low coverage; ModelFinder (Kalyaanamoorthy et al. 2017) was used to test the model and a maximum likelihood tree is generated using IQ-TREE (Nguyen et al. 2015) . The current version, 5.0.0, contains 7562 organisms: 4445 Bacteria, 168 Archaea, 447 Eukaryotes and 2502 Viruses.

A search is divided in multiple steps. The user has to first enter a search term (e.g. COX20), then select a species and indicate, in which taxonomic range the orthologous group should be searched in. When searching for COX20 and adding *H. sapiens* in EggNOG 5, the OG group ENOG502S3BD is found, containing 276 proteins found in 256 species, which indicates that inparalogs are included in OGs in EggNOG (Figure 12 a). The budding yeast COX20 ortholog is found in EggNOG v5, whereas it was not found in the previous version of the tool (4.5.1). Orthologs are listed with their identifiers from different databases, such as NCBI and Ensembl (Figure 12 b). The results include different tabs showing a taxonomic profile (Figure 12 c), GO terms, KEGG pathway, and conserved domains (Figure 12 d), a multiple alignment of all the orthologs (trimmed and untrimmed), as well as a phylogenetic tree that is decorated with functional annotation and conserved domains (Figure 12 e). All information for the protein family, including a Hidden Markov Model (HMM) consisting of all family members, is downloadable from the site. When searching for pyruvate carboxylase, 524 proteins are found in 367 species, the two inparalogs of *S. cerevisiae* were correctly found and are clustered together.

EggNOG provides an overview of the orthologous group tailless belongs to, which is called the steroid hormone mediated signaling pathway family. It also classifies the fine-grained, pairwise orthologs in this family. Tll, dsf and Hr51 all belong to the same EggNOG orthologous group. *Drosophila* has in total 14 members in this group, human 38. Tll itself is the pairwise ortholog of NR2E1, Hr51 of NR2E3. Finally, dsf has no human ortholog.

In conclusion, the use of EggNOG is quite easy and it can find remote orthologs and inparalogs. It moreover gives the complete overview of the orthologous group of tailless, which contains a number of paralogs in the different species, together with providing a detailed view on pairwise orthologs within this family. It applies a hierarchical procedure to cluster orthologs and offers a rich set of information and visualization of OGs. EggNOG is actively maintained and is available at : http://eggnog5.embl.de/#/app/home

PANTHER

PANTHER (Thomas et al. 2003) is part of the Gene Ontology Phylogenetic Annotation Project, led by the Gene Ontology consortium (Gaudet et al. 2011) . This acronym stands for Protein ANalysis THrough Evolutionary Relationships. It contains 142 genomes, 35 Bacteria, 8 Archaea, 99 Eukaryotes. Currently, version 15.0 is online, updated in February 2020. PANTHER's main goal is to provide high-confident functional annotations by classifying proteins according to their evolutionary history. Next to providing information on protein families and pathways, PANTHER also offers its own, reduced (i.e. slim) ontology for functional categorization. Since version 7, orthologs are annotated within PANTHER. PANTHER infers orthologs from family trees, based on pairs of genes who have diverged by a speciation event. Families are first separated based on the PANTHER HMM library. Multiple sequence alignments from these families are constructed using MAFFT (Katoh and Standley 2013) , which are then used for tree reconstruction using Giga (Thomas 2010). Not only one-to-one, but also one-to-many orthologs are inferred, reporting for instance inparalogs. In case of one-to-many relationships, PANTHER also reports the least diverged orthologs, which are believed to still have the same function.

When searching for human COX20 in 'genes and orthologs', a list of all orthologs is returned, with links to their entries in PANTHER (Figure 13 a) and gives ample information, such as IDs from other databases and alternate IDs (Figure 13 b), PANTHER families and subfamilies, the gene belongs to (Figure 13 c), PANTHER GO and GO slim annotations, as well as all orthologs of the gene (Figure 13 a). An interactive phylogenetic tree (full and reduced) can be displayed, which can be annotated with the full or trimmed multiple sequence alignment, including information on evolutionary events, such as deletions, insertions or mutations (Figure 13 d). Other types of information displayed include identifiers, family and sub-family associations, or panther IDs of functional annotations.

Next to searching in 'genes and orthologs', PANTHER families can also be searched. The term COX20 returns three families, whereby two correspond to the cytochrome c oxidase assembly protein COX20. Both are part of the same family and represent two different subfamilies, SF1 and SF2. The SF2 subfamily contains the human COX20 and three orthologs from primates whereas the SF1 subfamily contains the budding yeast protein and all other orthologs from a wide range of different species, including *Mus musculus*, *Danio rerio* or *Candida albicans*. The PANTHER family for COX20 (PTHR31586) contains a third sub-family, SF4, which is not named and which contains only *Macaca mulatta* and *Pan troglodytes*. An HMM of each family can be downloaded. Looking for pyruvate carboxylase (or PYC) in human, only one subfamily is found (PTHR43778). In fact, it is the only PANTHER family for this protein and consists of 73 genes. The two PYC genes of *S. cerevisiae* as well as human PC are in this PANTHER family.

PANTHER distinguishes the gene family and sub-families for the nuclear hormone receptors tll, dsf and Hr51. All three are part of the Nuclear Hormone Receptor (NHR) family, which includes 517 proteins from the supported 33 species and which are presented in the accompanying family tree at the PANTHER web-site. Tll, dsf and Hr51 are also part of their own sub-family, whereby tll groups with human NR2E1, Hr51 with NR2E3 and dsf is considered arthropod-specific. Gene counts of the NHR family, however differ between EggNOG and PANTHER: while EggNOG includes 14 Drosophila and 38 human members for this family, PANTHER only considers 8 Drosophila and 12 human proteins as part of the family.

Taken together, PANTHER is easy to handle, finds remote orthologs and gives information about inparalogs. Like EggNOG, PANTHER shows both, the entire family of nuclear hormone receptors for the tll family, as well as the sub-families with the consistent grouping of tll with NR2E1 and Hr51 with NR2E3. The division of organisms from the same orthologous group in sub-families can be confusing. PANTHER is available at : http://www.pantherdb.org/

Do-it-yourself: availability of search algorithms and orthology data from orthology resources

Most of the databases discussed provide their search algorithm and/or pre-calculated orthology data for download and local usage. This is specifically useful if users want to annotate their own genome, or place a large number of proteins in orthologous groups. On the other hand, pre-calculated orthologs can be useful for data mining or large-scale phylogenetic analyses.

Programmatic access and data download

Programmatic access, such as APIs (Application Programming Interfaces), is available for most resources. These interfaces allow users to download data from the database's website within their own pipelines and on a large-scale. All databases except for HCOP, P-POD and Inparanoid allow programmatic access to their data. Data from Homologene, HCOP or OMA additionally can be accessed directly via R.

All databases except for KEGG allow also download of their data, including multiple sequence alignments, HMM-libraries and phylogenetic trees, if available, which is useful for large-scale phylogenetic data mining. In order to avoid compatibility problems of formats, orthoXML (Schmitt et al. 2011) is used to store and compare orthology data from a wide range of databases, which is for instance offered by HCOP, InParanoid, OMA, OrthoMCL or PANTHER.

Availability of code

Several orthology inference tools are also downloadable as a stand-alone version. They can be used to identify orthologs and orthology groups of newly sequenced genomes assisting in the proper functional annotation of genes and proteins, phylogenetic profiling, or species tree reconstruction.

The pipeline used by OrthoDB can be downloaded for stand-alone usage. It represents the full pipeline used by the OrthoDB resource. After solving some dependencies for multithreading and boosting, the ORTHOPIPE and BRHCLUS software packages are easy to install and test locally. The stand-alone pipeline is available from the OrthoDB web-site (https://www.orthodb.org/?page=software).

An OMA standalone version is downloadable and usable as a command line tool (Altenhoff et al. 2019) . Installation and usage instructions are well-written and easy to follow; moreover parallelization instructions are provided to run larger OMA-jobs. The software package is very easy to install. OMA lists four major areas of application for its standalone version: species tree reconstruction, genome annotation, dynamics of genome evolution (making use of the HOG clusters) and finally phylogenetic profiling, looking for gene absences or duplications. Output is provided in OrthoXML format, as well as fasta and tabular files. OMA is available for download from the OMA web-site (https://omabrowser.org/standalone/#downloads), as well as from GitHub (https://github.com/DessimozLab/OmaStandalone/blob/master/OMA.drw).

OrthoMCL is available in a downloadable stand-alone version. Next to a local BLAST, it depends on a local installation of a relational database (MySQL or Oracle), for storing the orthologous pairs and orthology groups; as well as the MCL software for graph-based clustering. Installation instructions are available. The pipeline consists of individual perl scripts that need to be executed one after the other. Estimated run-time is given, which indicates the necessity of a larger compute cluster to run OrthoMCL locally. The software has only been tested for RedHat 5.8. The stand-alone version is available from the OrthoMCL web-site (https://orthomcl.org/common/downloads/software/v2.0/) MySQL, the MCL cluster software, as well as BLAST need to be downloaded and installed prior to installation of the pipeline. A new SQLite-dependent pipeline, which also contains a wrapper-script, is available from GitHub: (https://github.com/stajichlab/OrthoMCL).

Inparanoid is available as a stand-alone version for calculating pairwise orthologs, as well as orthologs among 3 organisms. It depends on NCBI-BLAST. Currently, InParanoid only supports the old version of BLAST and does not support the blast+ package, which is now standard. The user needs to either have a compatible version of BLAST available, or manipulate the inparanoid perl program to work with the newer blast+.

KEGG offers the BlastKOALA, GhostKOALA and KofamKOALA programs to automatically assign genome sequences to K numbers (KO assignment), however only as an online tool (Kanehisa et al. 2016b). The user can upload sequences for mapping to KEGG Orthology groups. Three different search algorithms are used: standard BLAST (BlastKOALA, https://www.kegg.jp/blastkoala/), or GHOSTX (GhostKOALA, https://www.kegg.jp/ghostkoala/), which is a fast homology search algorithm relying on query and database suffix arrays for seed matching; or in the newest version of the KEGG search tools, an HMM profile based search algorithm, HMMER3, (KofamKOALA (Aramaki et al. 2020) (https://www.genome.jp/tools/kofamkoala/)) is used. KofamKOALA searches against a pre-computed database of HMMs derived from KO families.

EggNOG offers the eggNOG mapper online (http://eggnog-mapper.embl.de/) and for download (https://github.com/eggnogdb/eggnog-mapper) to functionally annotate entire proteomes based on orthology. The stand-alone version can be easily cloned from GitHub and is easy to install. Sufficient documentation is provided, which is equally easy to follow. EggNOG mapper is also available online for annotating novel proteomes.

Finally, PANTHER provides the set of their tools for download at http://pantherdb.org/downloads/index.jsp, as well as on GitHub (https://github.com/pantherdb). It includes the PANTHER HMM scoring tool, which allows to compare a set of sequences, e.g. from a newly sequenced genome, against the entire PANTHER HMM library. The PantherScore tool depends on HMMER3, which needs to be installed independently. Amongst the tools provided is also the Java-based PAINT tree viewer program, as well as db-PAINT (from the GitHub repository) that allows functional annotation based on phylogenetic analysis. Installation instructions are given and easy to follow. Some essential information is missing; for example, it is unclear how the taxonomic ID file should be structured and how to retrieve it from NCBI.

Discussion

There exist many databases and tools that help identify orthologous genes or groups. We have presented the most commonly used and known available resources. We however do not claim that our list is exhaustive. We have tested the databases whether they can find remote orthologs and how they deal with paralogous genes, more precisely with inparalogs as well as with families with complex evolutionary histories. We found that all databases were able to handle inparalogs correctly. However, only few of them contained information on the remote ortholog we were searching for. Among those were OrthoDB, HCOP, KEGG, EggNOG and PANTHER. We would like to note at this point that while those resources were able to find COX20, we do not have any further data to support the claim that they contain all possible remote orthologs. The nuclear hormone receptor family tailless was particularly difficult to place for many resources. This is not surprising, as this family shows a large expansion in some taxonomic phyla, such as nematodes. While most of the databases were able to still correctly limit this group to the core members of *D. melanogaster* and *H. sapiens*, many of them contained putative false-positive members from *C. elegans*. Particularly TreeFam and P-POD fail to correctly classify this family. TreeFam for example contains proteins from *C. elegans* that are considered orthologous to other nuclear hormone receptor families in other resources: *C. elegans* unc-55 is for instance orthologous to NR2F1 from *H. sapiens*. It is furthermore noteworthy that two orthology databases, HomoloGene and InParanoid, define dsf as the ortholog of human NR2E1, instead of tll, even though it is not the best reciprocal hit, but only the second-best hit. This can be explained by the higher number of identical amino acids in local alignments between dsf and NR2E1 and the 239 amino acid long insertion in the middle of the dsf protein, which renders the RBH results ambiguous.

Different databases typically use different identifiers. Most accept gene names as a search item. Yet, ambiguous gene names such as the official gene symbol for pyruvate carboxylase, PC, find too many hits in the databases. Navigating search results can therefore be problematic. A text-based search for the full name, or using an identifier accepted by the database resolved this problem.

Orthology databases differ in the availability of additional annotation provided for orthologous groups. In this respect, databases embedded in genome resources have an advantage, as the entire information collected on genes is easily available. These include for instance HomoloGene, PANTHER, or KEGG.

Some databases do not contain up-to-date information. This means in most cases that no new species were integrated in the database. While we do not see this as a reason for not making use of a resource, it indicates that curation might have been neglected for these databases.

Among the databases tested, we found that OrthoDB was one of the databases with the largest number of available organisms. It also had the most complete set of orthologs, including remotely conserved orthologs. It has linked all entries to several databases, provides domain annotation, as well as functional annotation of orthologous groups. Moreover, the NCBI gene resource meanwhile relies for orthology information next to HomoloGene also on OrthoDB. We see the hierarchical treatment of orthologous groups employed by OrthoDB, but also by other resources like EggNOG as an advantage. Functions are more likely retained in closely related species and thus, making use of a more fine-grained, taxonomical clustering will result in more accurate functional annotation transfer.

The availability of code, as well as web-based services are required for annotation of newly sequenced genomes. Many resources allow users to either download their code locally; or provide web-based services for whole-genome annotations. While we have not tested each available tool locally, we found that they are easy enough to install and usage instructions are easy to follow. To run such computationally heavy tools locally can however exceed dramatically the resources available to a user, both in compute time, as well as in storage space. When considering orthology-based functional assignment, this should be kept in mind. Using web-based services like EggNOG, KEGG or PANTHER provides a viable solution to this problem.

Finally, we want to stress the importance of adding newly sequenced genomes to orthology search pipelines. First of all, novel model organisms arise rapidly, for instance to address questions in evolutionary developmental biology. Providing a good functional annotation of sequences by knowledge transfer from orthologs will help advance scientific discovery in non-standard model organisms. Second, adding new species will lead to a better coverage of search space to find orthologs. This will ultimately also help in discovering remote orthologies and in gaining a better understanding of the evolution of genes and pathways.

Acknowledgements

We would like to thank the support staff from the NCBI for help with data retrieval and for providing information; Maria-Mandela Prunster, Amélie Vernale and Friedhelm Pfeiffer for critical reading of the manuscript. This work was supported by the CNRS and by an MENRT thesis grant from the French Ministry of Research awarded to Paul de Boissier.

References

Altenhoff AM, Glover NM, Train C-M, et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 46:D477–D485. doi: 10.1093/nar/gkx1019

Altenhoff AM, Levy J, Zarowiecki M, et al (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res 29:1152–1163. doi: 10.1101/gr.243212.118

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8:e1002514. doi: 10.1371/journal.pcbi.1002514

Altenhoff AM, Škunca N, Glover N, et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–9. doi: 10.1093/nar/gku1158

Altschul SF, Gish W (1996) Local alignment statistics. Meth Enzymol 266:460–480. doi: 10.1016/s0076-6879(96)66029-7

Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi: 10.1093/nar/25.17.3389

Aramaki T, Blanc-Mathieu R, Endo H, et al (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36:2251–2252. doi: 10.1093/bioinformatics/btz859

Arnold R, Goldenberg F, Mewes H-W, Rattei T (2014) SIMAP--the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. Nucleic Acids Res 42:D279–84. doi: 10.1093/nar/gkt970

Brown D, Sjölander K (2006) Functional classification using phylogenomic inference. PLoS Comput Biol 2:e77. doi: 10.1371/journal.pcbi.0020077

Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–8. doi: 10.1093/nar/gkj123

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7:429–447. doi: 10.1089/106652700750050871

Cunningham F, Achuthan P, Akanni W, et al (2019) Ensembl 2019. Nucleic Acids Res 47:D745–D751. doi: 10.1093/nar/gky1113

El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. doi: 10.1093/nar/gky995

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584. doi: 10.1093/nar/30.7.1575

Eyre TA, Wright MW, Lush MJ, Bruford EA (2007) HCOP: a searchable database of human orthology predictions. Brief Bioinformatics 8:2–5. doi: 10.1093/bib/bbl030

Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Res 40:D136–43. doi: 10.1093/nar/gkr1178

Fischer S, Brunk BP, Chen F, et al (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6:Unit 6.12.1–19. doi: 10.1002/0471250953.bi0612s35

Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinformatics 12:449–462. doi: 10.1093/bib/bbr042

Guindon S, Dufayard J-F, Lefort V, et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. doi: 10.1093/sysbio/syq010

Heinicke S, Livstone MS, Lu C, et al (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PLoS ONE 2:e766. doi: 10.1371/journal.pone.0000766

Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635–1638. doi: 10.1093/molbev/msw046

Huerta-Cepas J, Szklarczyk D, Heller D, et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. doi: 10.1093/nar/gky1085

Jaccard P (1912) THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytologist 11:37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Jensen LJ, Julien P, Kuhn M, et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250–4. doi: 10.1093/nar/gkm796

Kalyaanamoorthy S, Minh BQ, Wong TKF, et al (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. doi: 10.1038/nmeth.4285

Kanehisa M, Furumichi M, Tanabe M, et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361. doi: 10.1093/nar/gkw1092

Kanehisa M, Goto S, Furumichi M, et al (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38:D355–60. doi: 10.1093/nar/gkp896

Kanehisa M, Goto S, Sato Y, et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–205. doi: 10.1093/nar/gkt1076

Kanehisa M, Sato Y, Kawashima M, et al (2016a) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–62. doi: 10.1093/nar/gkv1070

Kanehisa M, Sato Y, Morishima K (2016b) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 428:726–731. doi: 10.1016/j.jmb.2015.11.006

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. doi: 10.1093/molbev/mst010

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338. doi: 10.1146/annurev.genet.39.073003.114725

Kriventseva EV, Kuznetsov D, Tegenfeldt F, et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 47:D807–D811. doi: 10.1093/nar/gky1053

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res 36:D271–5. doi: 10.1093/nar/gkm845

Li H, Coghlan A, Ruan J, et al (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–80. doi: 10.1093/nar/gkj118

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. doi: 10.1101/gr.1224503

Livstone MS, Oughtred R, Heinicke S, et al (2011) Inferring protein function from homology using the Princeton Protein Orthology Database (P-POD). Curr Protoc Bioinformatics Chapter 6:Unit 6.11. doi: 10.1002/0471250953.bi0611s33

Lu S, Wang J, Chitsaz F, et al (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. doi: 10.1093/nar/gkz991

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7–19. doi: 10.1093/nar/gkv1290

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13. doi: 10.1093/nar/gkx1095

Nguyen L-T, Schmidt HA, Haeseler von A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. doi: 10.1093/molbev/msu300

O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33:D476–80. doi: 10.1093/nar/gki107

Pronk JT, Yde Steensma H, Van Dijken JP (1996) Pyruvate metabolism in Saccharomyces cerevisiae. Yeast 12:1607–1633. doi: 10.1002/(sici)1097-0061(199612)12:16<1607::aid-yea70>3.0.co;2-4

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. doi: 10.1016/s0168-9525(00)02024-2

Ruan J, Li H, Chen Z, et al (2008) TreeFam: 2008 Update. Nucleic Acids Res 36:D735–40. doi: 10.1093/nar/gkm1005

Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief Bioinformatics 12:485–488. doi: 10.1093/bib/bbr025

Schreiber F, Patricio M, Muffato M, et al (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res 42:D922–5. doi: 10.1093/nar/gkt1055

Sievers F, Wilm A, Dineen D, et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi: 10.1038/msb.2011.75

Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234–9. doi: 10.1093/nar/gku1203

Steinegger M, Meier M, Mirdita M, et al (2019) HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473–15. doi: 10.1186/s12859-019-3019-7

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. doi: 10.1038/nbt.3988

Szklarczyk R, Wanschers BFJ, Nijtmans LG, et al (2013) A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. Hum Mol Genet 22:656–667. doi: 10.1093/hmg/dds473

Tatusov RL, Fedorova ND, Jackson JD, et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41–14. doi: 10.1186/1471-2105-4-41

Thomas PD (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics 11:312–19. doi: 10.1186/1471-2105-11-312

Thomas PD, Campbell MJ, Kejariwal A, et al (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141. doi: 10.1101/gr.772403

Walter F (1989) R. F. Doolittle, Of URFS and ORFS — a Primer on How to Analyze Derived Amino Acid Sequences. VII + 103 S., 24 Abb., 14 Tab. Mill Valley 1986. University Science Books. ISBN: 0-935702-54-7. Journal of Basic Microbiology 29:246–246. doi: 10.1002/jobm.3620290411

Wolf YI, Koonin EV (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol 4:1286–1294. doi: 10.1093/gbe/evs100

Zahn-Zabal M, Dessimoz C, Glover NM (2020) Identifying orthologs with OMA: A primer. F1000Res 9:27. doi: 10.12688/f1000research.21508.1

Tables

Table 1 : Information on the proteins used for testing orthology resources

Organism	Gene Name	Protein sequence ID	Gene sequence ID	Gene ID	Ensembl ID / locus tag
H. sapiens	COX20 (FAM36A)	NP_001299800	NM_001312871	116228	ENSG00000203667
S. cerevisiae	COX20	NP_010517	NM_001180539	851817	YDR231C
H. sapiens	PC	NP_000911	NM_000920	5091	ENSG00000173599
S. cerevisiae	PYC1	NP_011453	NM_001180927	852818	YGL062W
S. cerevisiae	PYC2	NP_009777	NM_001178566	852519	YBR218C
D. melanogaster	tll	NP_524596	NM_079857	43656	CG1378
D. melanogaster	dsf	NP_477140	NM_057792	33823	CG9019
D. melanogaster	Hr51	NP_611032	NM_137188	36702	CG16801
H. sapiens	NR2E1	NP_001273031	NM_001286102	7101	ENSG00000112333
H. sapiens	NR2E3	NP_057430	NM_016346	10002	ENSG00000278570

Table 2 : Summary of the tools studied. Given information relies on the version available in January 2020.

Tools	Method	Number of species	(B)acteria / (A)rchaea / (E)ukaryota	Remote orthologs	Inparalogs	Programmatic access	Data download	References
OrthoDB	RBH + clustering	13772	B / A / E	✓	✓	✓	✓	Kriventseva Evgenia V et al. 2019
HomoloGene	RBH + Tree	21	E	✗	✓	✓	✓	N.C.B.I Resource Coordinators 2013
TreeFam	Tree	109	E	✗	✓	✓	✓	Ruan J et al. 2007
HCOP	Combination of databases	19	E (human-centric)	✓	✓	✗	✓	Eyre Tina A et al. 2007
OMA	Smith and Waterman + clustering	2288	B / A / E	✗	✓	✓	✓	Altenhoff Adrian M et al. 2018 Altenhoff A M et al. 2015

OrthoMCL	Markov Cluster algorithm + graph flow theory	150	B / A / E	✗	✓	✓	✓	Li Li et al. 2003 Chen Feng et al. 2006
P-POD	Combination of databases + Tree	12	E	✗	✓	✗	✓	Heinicke Sven et al. 2007
InParanoid	RBH + clustering	273	B / A / E	✗	✓	✗	✓	O'Brien Kevin P et al. 2005 Sonnhammer E L L et al. 2015
KEGG Orthology	KOALA + Manual curation	>4000	B / A / E	✓	✓	✓	✗	Kanehisa Minoru et al. 2016 Kanehisa Minoru et al. 2014
EggNOG	Smith and Waterman + clustering + tree + HMM	7562	B / A / E	✓	✓	✓	✓	Huerta-Cepas Jaime et al. 2019
PANTHER	HMM + tree	142	B / A / E	✓	✓	✓	✓	Thomas Paul D. et al 2003

Figure legends

Figure 1 : Growth of sequence databases and completely sequenced genomes in the different kingdoms. (a) Cumulative number of sequences found on the NCBI GenBank and for Whole Genome Sequences (WGS) since 1985. (b) Number of sequenced genomes available at NCBI.

Figure 2 : Schematic representation of homologous relationships. Orthologs and paralogs are produced respectively by speciation and duplication events. Inparalogs and outparalogs are paralogs produced after or before a speciation event. (a) Inparalogs are the result of a direct duplication event. Two inparalogs are therefore more closely related to each other than to any other gene in another organism. (b) Outparalogs on the other hand are the result of a duplication event followed by a speciation event. Colored circles represent species, blue boxes represent a gene and small white circles represent mutations.

Figure 3 : Sequences chosen for testing different orthology resources. (a) Pairwise alignment of COX20 from *H. sapiens* and *S. cerevisiae*. Similar residues are highlighted in yellow. The two proteins share below 30% of sequence similarity and only 14% of sequence identity, which makes them remote orthologs. (b) Tree representation of the relationship between the pyruvate carboxylases. Inparalogs are more similar to each other than their ortholog in *H. sapiens*. (c) Phylogenetic tree of the tailless family, showing the relationship of the three proteins tailless (tll), dissatisfied (dsf) and hormone receptor 51 (Hr51) with the human family members NR2E1 and NR2E3. Proteins were aligned using mafft (Katoh and Standley 2013), the tree was calculated using IQ-TREE (Nguyen et al. 2015) with the -s option and 1000 iterations.

Figure 4 : COX20 OrthoDB group at Eukaryota level. (a) Interactive group hierarchy split in five different subgroups. (b) Annotations of the orthologous group, including GO terms, InterPro domains and evolutionary information. (c) List of orthologs classified by organisms in a taxonomical, tree-like structure. Taxonomical levels can be displayed or hidden by making use of the arrow. For human COX20, all associated information is shown; identifiers are linked to their respective database of origin.

Figure 5 : Results of the HomoloGene database. (a) COX20 is conserved in Euteloestomi., which includes *H. sapiens.* Protein domains and sequence lengths are displayed. Links are provided to the gene and the protein entries of the NCBI database. (b) HomoloGene allows to see alignments and launch a pairwise BLAST, or to retrieve pairwise alignment scores. A list of publications linked to the family members is given. (c) Multiple sequence alignment of the protein family. (d) Pyruvate carboxylase results. Diverse domains are colored differently and link to their respective CDD entry. The results page was restricted to show only the results for *H. sapiens* and *S. cerevisiae*. (e) BLASTp alignment of NR2E1 with tailless (tll) and dissatisfaction (dsf). Dsf shows a higher number of identical amino acids in the local alignment, while being the second-best hit found in *D. melanogaster*, when searched with human NR2E1.

Figure 6 : Treefam search showing the model trees for COX20 and PC. (a) Model tree of the COX20 family in TreeFam. Different tabs allow to access different information, such as the sequences and the Wikipedia page. A download page is available to retrieve information associated with the protein family. (b) Reduced phylogenetic tree of the PC family, indicating gene duplication (red arrows) and speciation (green dots) events. (c) TreeFam tree of the tailless family. Hr83, nhr-14, unc-55 as well as several other *C. elegans* proteins are false-positive members of this group.

Figure 7 : The HCOP database search page and results. (a) Input options for a HCOP search, presenting all the species and databases from which the information can be extracted. (b) Results of HCOP for COX20 for *Drosophila melanogaster* and *S. cerevisiae*. The results provided are derived from five different databases for the fruit fly; the orthologous budding yeast COX20 protein could only retrieved from the PANTHER resource.

Figure 8 : OMA orthology database results pages. (a) Tabular listing of 1:1 orthologs for the human COX20 protein. (b) The local synteny (four genes upward and downward) is displayed by small colored boxes with the direction of transcription. (c) The hierarchical orthologous groups represent the orthologs as a tree with the number of copies per species. The Chordata node is displayed in red. (d) the tree can be decorated with information on the gene length or the GC content (here shown for human COX20). (e) Multiple sequence alignment of the OMA group. A logo is extracted from it, in which the size of the letter for an amino acid is representative for its conservation. (f) OMA groups in which the COX20 proteins from different species are included. (g) GO terms associated with each ortholog.

Figure 9 : OrthoMCL DB results. (a) Results for a search with human COX20. Most orthologs are found in Metazoans. (b) Graph representation of the COX20 ortholog groups. Individual nodes represent proteins, edges between them represent their orthologous relationship. The graph can be manipulated by choosing different parameters in the control panel on the left-hand side control panel.

Figure 10 : P-POD results for pyruvate carboxylase. (a) Using the Jaccard algorithm, four proteins are found for *S. cerevisiae* and seven in *H. sapiens,* while OrthoMCL based clustering retrieves one ortholog in human and the two inparalogs in budding yeast. (b) Tree representation of the P-POD results. Duplication events are displayed by red squares.

Figure 11 : Simplistic results of the Inparanoid and KEGG orthology database. (a) InParanoid results for COX20. As *S. cerevisiae* was not found, we show an example with *D. melanogaster*. (b) InParanoid results for pyruvate carboxylase. The inparalogs are grouped with their ortholog in *H. sapiens*. (c) KEGG results for COX20. Next to the basic information on the protein, the main information provided comes from the Brite annotations. The list of orthologs is shown in the main box, as are potential articles associated with an ortholog group. Links to all other resources from KEGG are given in the ‘All links’ box.

Figure 12 : EggNOG information provided for ortholog groups, here for COX20. (a) EggNOG group for COX20 at the Eukaryota level with a brief functional description, as well as the number of proteins and organisms found (top boxes). (b) An exhaustive list of orthologs is given, linking to different source databases to retrieve the sequence (restricted for illustration purposes). (c) the taxonomic profile of COX20. The profile is interactive and the user can

browse at different levels of the taxonomic hierarchy. Here shown for *S. cerevisiae* (levels passed are shown in this case in dark pink). (d) GO terms associated with the COX20 group, providing information on the evidence of a GO term. (e) Phylogenetic tree that is decorated with aligned segments of the ortholog group. (f) EggNOG results for PC, with the two identified inparalogs, PYC1 and PYC2, in *S. cerevisiae*.

Figure 13 : PANTHER information on the COX20 family. (a) List of orthologs of COX20 with links to their entries in PANTHER. (b) Panther gene information about COX20, linking via identifiers to different databases. (c) Classification of the gene in PANTHER and the PANTHER slim ontology. (d) Interactive phylogenetic tree, annotated with the multiple sequence alignment used to calculate the tree, including information on evolutionary events, such as deletions, insertions or mutations.

Figures

Figure 1

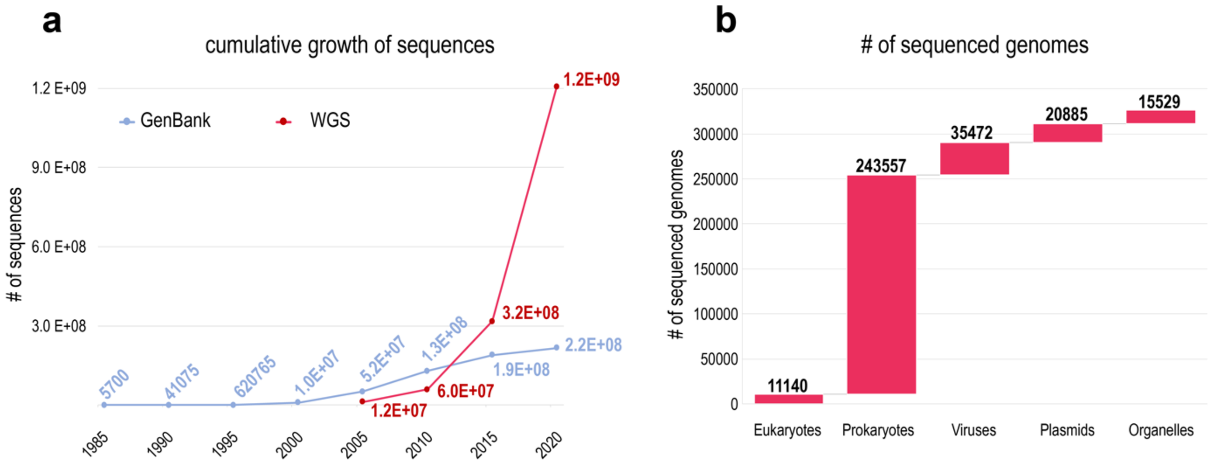


Figure 2

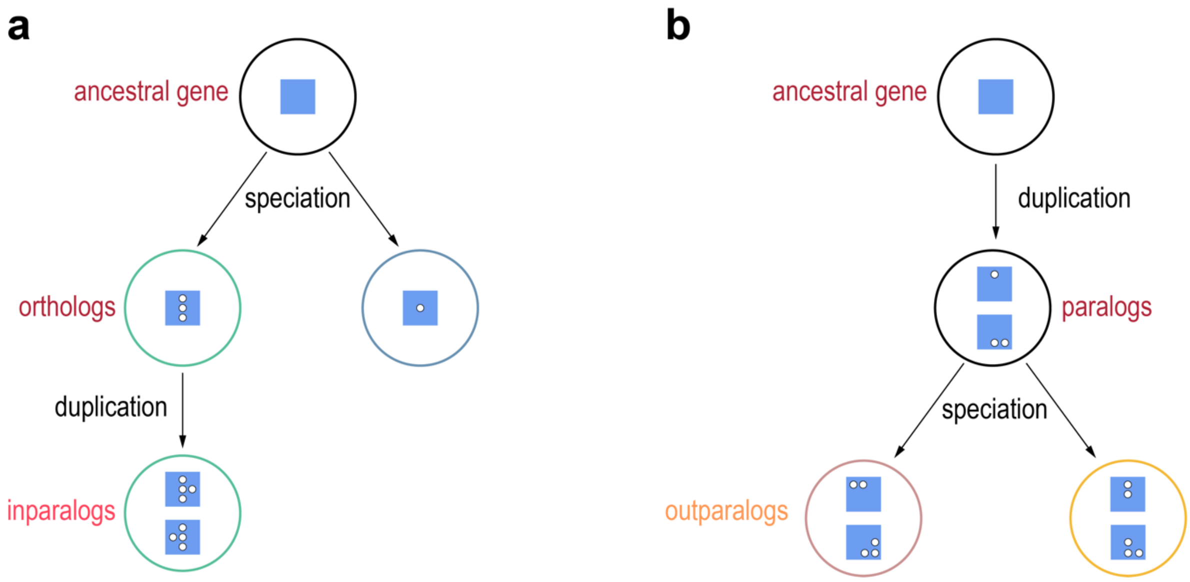


Figure 3

a

```
H.sapiens         ----------------------------------------MAAPP---------EPGEPE
S.cerevisiae      MRWWPWSNQTEDQKQQQQPQGKADGDRVLTNYSRGQKILLEDTPPKFADDLSNSQLAKKQ

H.sapiens         ERKSLKL---------LGFLDVENTPCARHSILYG-----SLGSVVAGFGHFLFTSRIRR
S.cerevisiae      ERATLKEAWDSIRWSDFSLQKLTSIPCFRDAGMLGFSSMFLMGSII-----FIYHKSPTK

H.sapiens         SCDVGVGGFILVTLGCWFHCRYNYAKQRIQERIAREEIKKK---ILYEGTHLDP------
S.cerevisiae      ATNWAMSSLILGSIVGWEQCRLKRQKSFQIAQLAKETVAKKEKPMLHNVPH-DPSLPGQW

H.sapiens         -----ERKHNGSSSN----------------
S.cerevisiae      EAAKNEKQSQFEQSNQNLSQASSEKKWYKFW
```

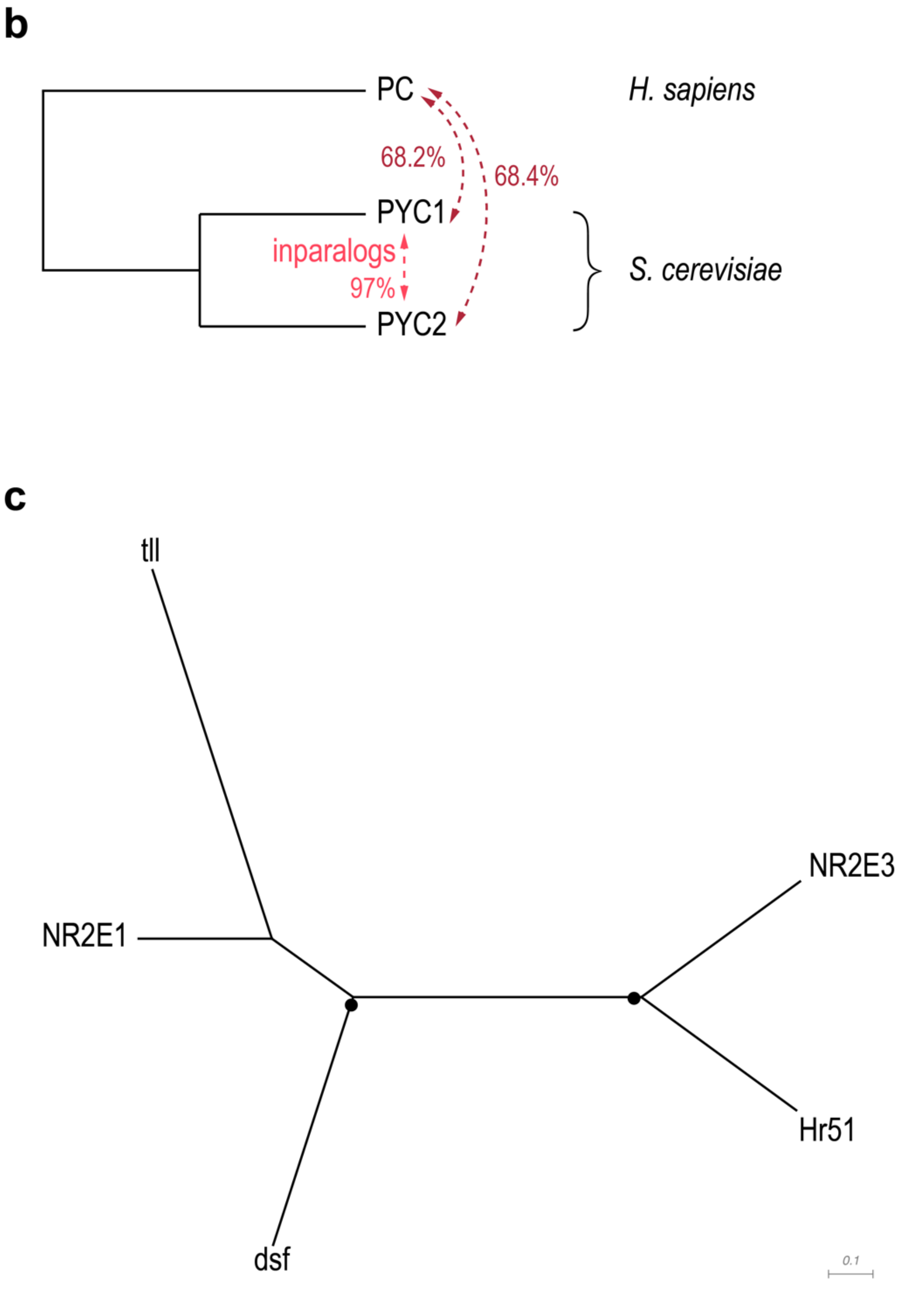
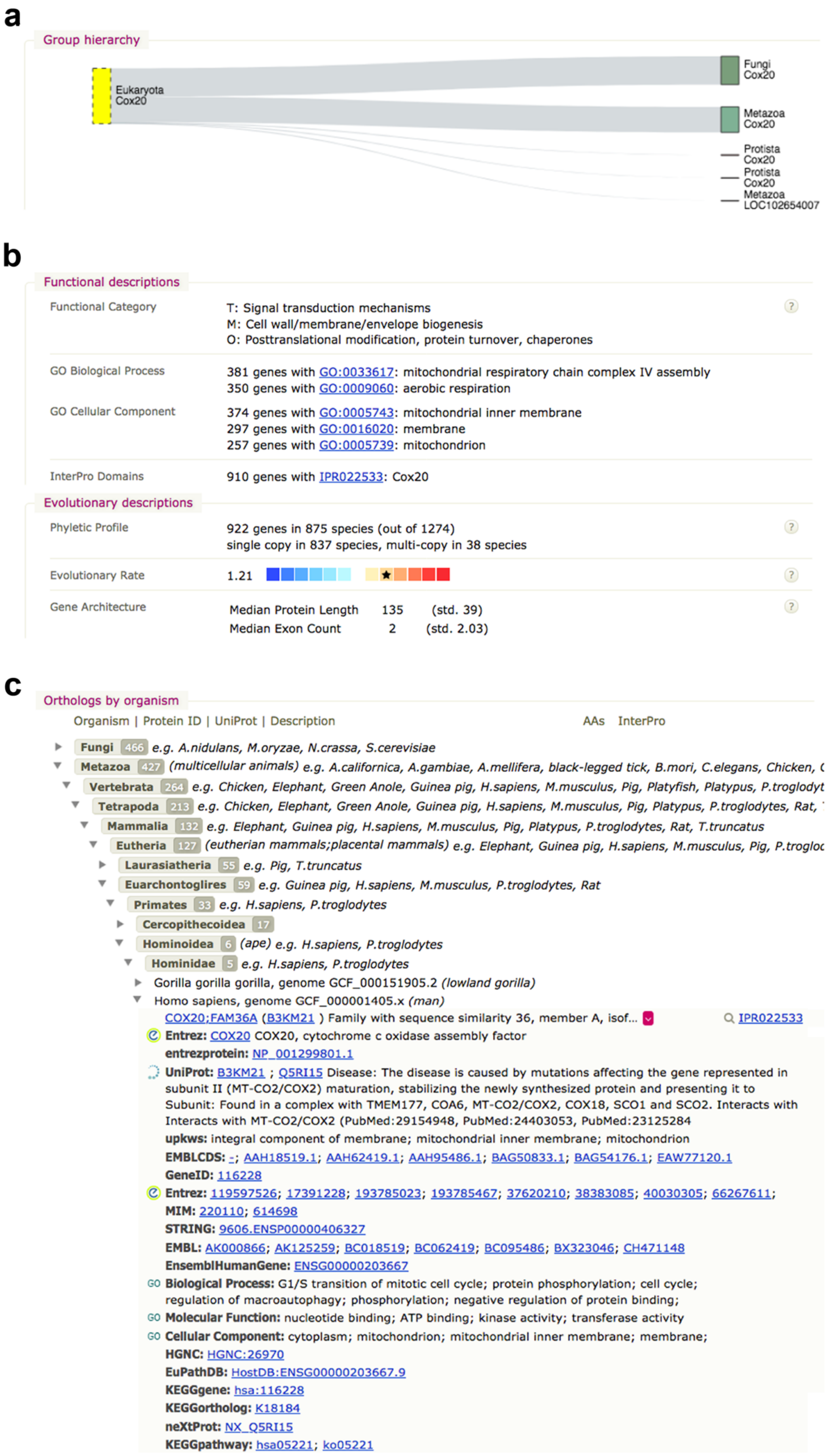


Figure 4



Functional descriptions

Functional Category

T: Signal transduction mechanisms
M: Cell wall/membrane/envelope biogenesis
O: Posttranslational modification, protein turnover, chaperones

GO Biological Process

381 genes with GO:0033617: mitochondrial respiratory chain complex IV assembly
350 genes with GO:0009060: aerobic respiration

GO Cellular Component

374 genes with GO:0005743: mitochondrial inner membrane
297 genes with GO:0016020: membrane
257 genes with GO:0005739: mitochondrion

InterPro Domains

910 genes with IPR022533: Cox20

Evolutionary descriptions

Phyletic Profile

922 genes in 875 species (out of 1274)
single copy in 837 species, multi-copy in 38 species

Evolutionary Rate

1.21 ★

Gene Architecture

Median Protein Length 135 (std. 39)
Median Exon Count 2 (std. 2.03)

Orthologs by organism

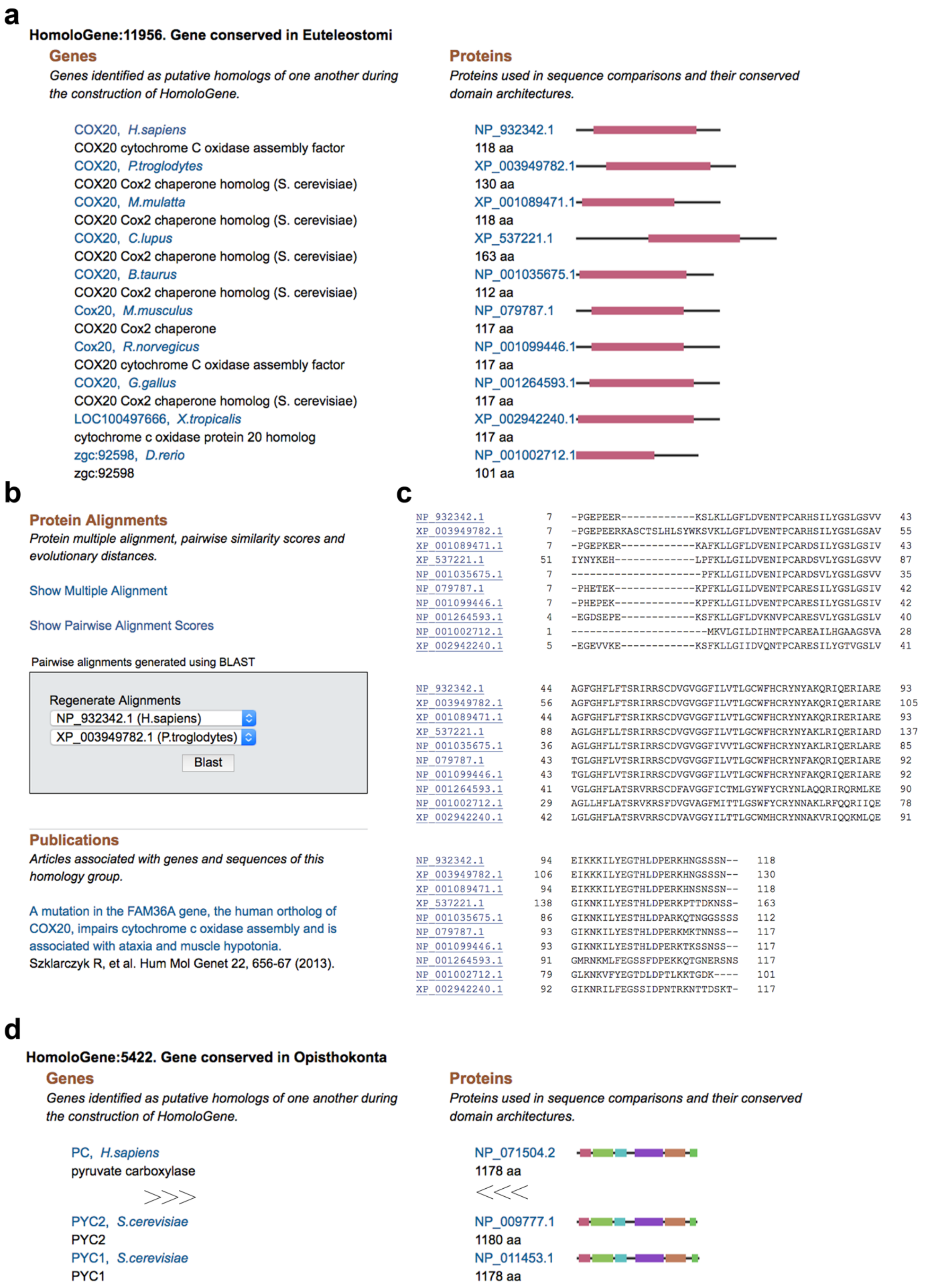
Organism | Protein ID | UniProt | Description AAs InterPro

▶ **Fungi** 466 *e.g. A.nidulans, M.oryzae, N.crassa, S.cerevisiae*
▼ **Metazoa** 427 *(multicellular animals) e.g. A.californica, A.gambiae, A.mellifera, black-legged tick, B.mori, C.elegans, Chicken,*
▼ **Vertebrata** 264 *e.g. Chicken, Elephant, Green Anole, Guinea pig, H.sapiens, M.musculus, Pig, Platyfish, Platypus, P.troglodyt*
▼ **Tetrapoda** 213 *e.g. Chicken, Elephant, Green Anole, Guinea pig, H.sapiens, M.musculus, Pig, Platypus, P.troglodytes, Rat,*
▼ **Mammalia** 132 *e.g. Elephant, Guinea pig, H.sapiens, M.musculus, Pig, Platypus, P.troglodytes, Rat, T.truncatus*
▼ **Eutheria** 127 *(eutherian mammals;placental mammals) e.g. Elephant, Guinea pig, H.sapiens, M.musculus, Pig, P.trogloc*
▶ **Laurasiatheria** 55 *e.g. Pig, T.truncatus*
▼ **Euarchontoglires** 59 *e.g. Guinea pig, H.sapiens, M.musculus, P.troglodytes, Rat*
▼ **Primates** 33 *e.g. H.sapiens, P.troglodytes*
▶ **Cercopithecoidea** 17
▼ **Hominoidea** 6 *(ape) e.g. H.sapiens, P.troglodytes*
▼ **Hominidae** 5 *e.g. H.sapiens, P.troglodytes*
▶ Gorilla gorilla gorilla, genome GCF_000151905.2 *(lowland gorilla)*
▼ Homo sapiens, genome GCF_000001405.x *(man)*

COX20;FAM36A (B3KM21) Family with sequence similarity 36, member A, isof... IPR022533

Entrez: COX20 COX20, cytochrome c oxidase assembly factor
entrezprotein: NP_001299801.1
UniProt: B3KM21 ; Q5RI15 Disease: The disease is caused by mutations affecting the gene represented in subunit II (MT-CO2/COX2) maturation, stabilizing the newly synthesized protein and presenting it to Subunit: Found in a complex with TMEM177, COA6, MT-CO2/COX2, COX18, SCO1 and SCO2. Interacts with Interacts with MT-CO2/COX2 (PubMed:29154948, PubMed:24403053, PubMed:23125284
upkws: integral component of membrane; mitochondrial inner membrane; mitochondrion
EMBLCDS: -; AAH18519.1; AAH62419.1; AAH95486.1; BAG50833.1; BAG54176.1; EAW77120.1
GeneID: 116228
Entrez: 119597526; 17391228; 193785023; 193785467; 37620210; 38383085; 40030305; 66267611;
MIM: 220110; 614698
STRING: 9606.ENSP00000406327
EMBL: AK000866; AK125259; BC018519; BC062419; BC095486; BX323046; CH471148
EnsemblHumanGene: ENSG00000203667
GO **Biological Process:** G1/S transition of mitotic cell cycle; protein phosphorylation; cell cycle; regulation of macroautophagy; phosphorylation; negative regulation of protein binding;
GO **Molecular Function:** nucleotide binding; ATP binding; kinase activity; transferase activity
GO **Cellular Component:** cytoplasm; mitochondrion; mitochondrial inner membrane; membrane;
HGNC: HGNC:26970
EuPathDB: HostDB:ENSG00000203667.9
KEGGgene: hsa:116228
KEGGortholog: K18184
neXtProt: NX_Q5RI15
KEGGpathway: hsa05221; ko05221

Figure 5



HomoloGene:11956. Gene conserved in Euteleostomi

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

COX20, *H.sapiens*
COX20 cytochrome C oxidase assembly factor
COX20, *P.troglodytes*
COX20 Cox2 chaperone homolog (S. cerevisiae)
COX20, *M.mulatta*
COX20 Cox2 chaperone homolog (S. cerevisiae)
COX20, *C.lupus*
COX20 Cox2 chaperone homolog (S. cerevisiae)
COX20, *B.taurus*
COX20 Cox2 chaperone homolog (S. cerevisiae)
Cox20, *M.musculus*
COX20 Cox2 chaperone
Cox20, *R.norvegicus*
COX20 cytochrome C oxidase assembly factor
COX20, *G.gallus*
COX20 Cox2 chaperone homolog (S. cerevisiae)
LOC100497666, *X.tropicalis*
cytochrome c oxidase protein 20 homolog
zgc:92598, *D.rerio*
zgc:92598

Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

NP_932342.1
118 aa
XP_003949782.1
130 aa
XP_001089471.1
118 aa
XP_537221.1
163 aa
NP_001035675.1
112 aa
NP_079787.1
117 aa
NP_001099446.1
117 aa
NP_001264593.1
117 aa
XP_002942240.1
117 aa
NP_001002712.1
101 aa

Protein Alignments

Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Show Multiple Alignment

Show Pairwise Alignment Scores

Pairwise alignments generated using BLAST

Regenerate Alignments
NP_932342.1 (H.sapiens)
XP_003949782.1 (P.troglodytes)

Publications

Articles associated with genes and sequences of this homology group.

A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia.
Szklarczyk R, et al. Hum Mol Genet 22, 656-67 (2013).

```
NP_932342.1         7   -PGEPEER------------KSLKLLGFLDVENTPCARHSILYGSLGSVV   43
XP_003949782.1      7   -PGEPEERKASCTSLHLSYWKSVKLLGFLDVENTPCARHSILYGSLGSAV   55
XP_001089471.1      7   -PGEPKER------------KAFKLLGFLDVENIPCARDSILYGSLGSIV   43
XP_537221.1        51   IYNYKEH-------------LPFKLLGILDVENIPCARDSVLYGSLGSVV   87
NP_001035675.1      7   ---------------------PFKLLGILDVENIPCARDSVLYGSLGSVV   35
NP_079787.1         7   -PHETEK-------------KPFKLLGILDVENTPCARESILYGSLGSIV   42
NP_001099446.1      7   -PHEPEK-------------KPFKLLGILDVENTPCARESILYGSLGSIV   42
NP_001264593.1      4   -EGDSEPE------------KSFKLLGFLDVKNVPCARESVLYGSLGSLV   40
NP_001002712.1      1   ----------------------MKVLGILDIHNTPCAREAILHGAAGSVA   28
XP_002942240.1      5   -EGEVVKE------------KSFKLLGIIDVQNTPCARESILYGTVGSLV   41


NP_932342.1        44   AGFGHFLFTSRIRRSCDVGVGGFILVTLGCWFHCRYNYAKQRIQERIARE   93
XP_003949782.1     56   AGFGHFLFTSRIRRSCDVGVGGFILVTLGCWFHCRYNYAKQRIQERIARE   105
XP_001089471.1     44   AGFGHFLFTSRIKRSCDVGVGGFILVTLGCWFHCRYNYAKQRIRERIARE   93
XP_537221.1        88   AGLGHFLLTSRIRRSCDVGVGGFILVTLGCWFHCRYNYAKLRIQERIARD   137
NP_001035675.1     36   AGLGHFLLTSRIRRSCDVGVGGFIVVTLGCWFHCRYNYAKLRIQERLARE   85
NP_079787.1        43   TGLGHFLVTSRIRRSCDVGVGGFILVTLGCWFHCRYNFAKQRIQERIARE   92
NP_001099446.1     43   TGLGHFLVTSRIRRSCDVGVGGFILVTLGCWFHCRYNFAKQRIQERIARE   92
NP_001264593.1     41   VGLGHFLATSRVRRSCDFAVGGFICTMLGYWFYCRYNLAQQRIRQRMLKE   90
NP_001002712.1     29   AGLLHFLATSRVKRSFDVGVAGFMITTLGSWFYCRYNNAKLRFQQRIIQE   78
XP_002942240.1     42   LGLGHFLATSRVRRSCDVAVGGYILTTLGCWMHCRYNNAKVRIQQKMLQE   91


NP_932342.1        94   EIKKKILYEGTHLDPERKHNGSSSN--   118
XP_003949782.1    106   EIKKKILYEGTHLDPERKHNGSSSN--   130
XP_001089471.1     94   EIKKKILYEGTHLDPERKHNSNSSN--   118
XP_537221.1       138   GIKNKILYESTHLDPERKPTTDKNSS-   163
NP_001035675.1     86   GIKNKILYESTHLDPARKQTNGGSSSS   112
NP_079787.1        93   GIKNKILYESTHLDPERKMKTNNSS--   117
NP_001099446.1     93   GIKNKILYESTHLDPERKTKSSNSS--   117
NP_001264593.1     91   GMRNKMLFEGSSFDPEKKQTGNERSNS   117
NP_001002712.1     79   GLKNKVFYEGTDLDPTLKKTGDK----   101
XP_002942240.1     92   GIKNRILFEGSSIDPNTRKNTTDSKT-   117
```

HomoloGene:5422. Gene conserved in Opisthokonta

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

PC, *H.sapiens*
pyruvate carboxylase

PYC2, *S.cerevisiae*
PYC2
PYC1, *S.cerevisiae*
PYC1

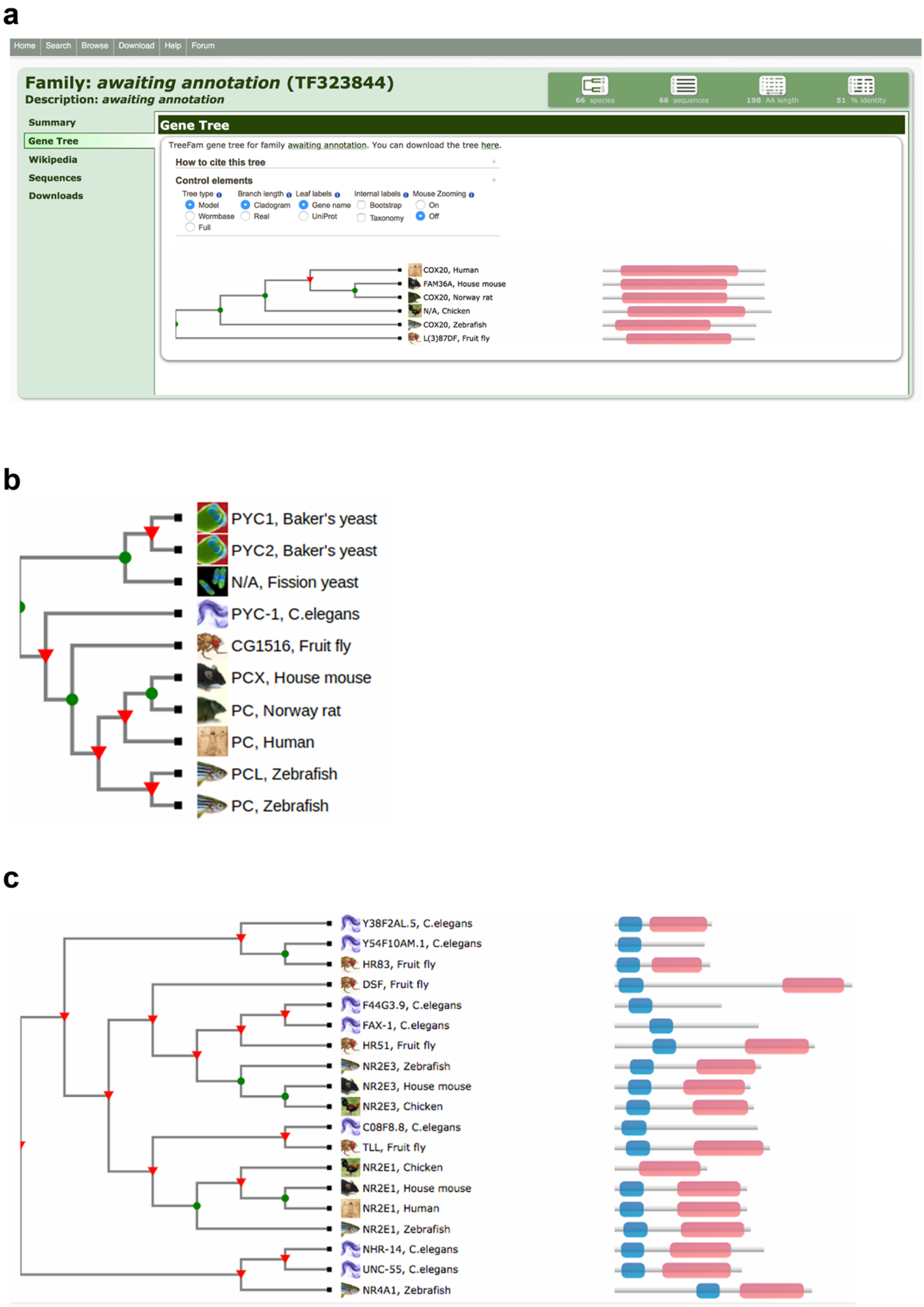
Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

NP_071504.2
1178 aa

NP_009777.1
1180 aa
NP_011453.1
1178 aa

Figure 6



Home | Search | Browse | Download | Help | Forum

Family: *awaiting annotation* (TF323844)
Description: *awaiting annotation*

66 species | 66 sequences | 198 AA length | 51 % identity

Summary
Gene Tree
Wikipedia
Sequences
Downloads

Gene Tree

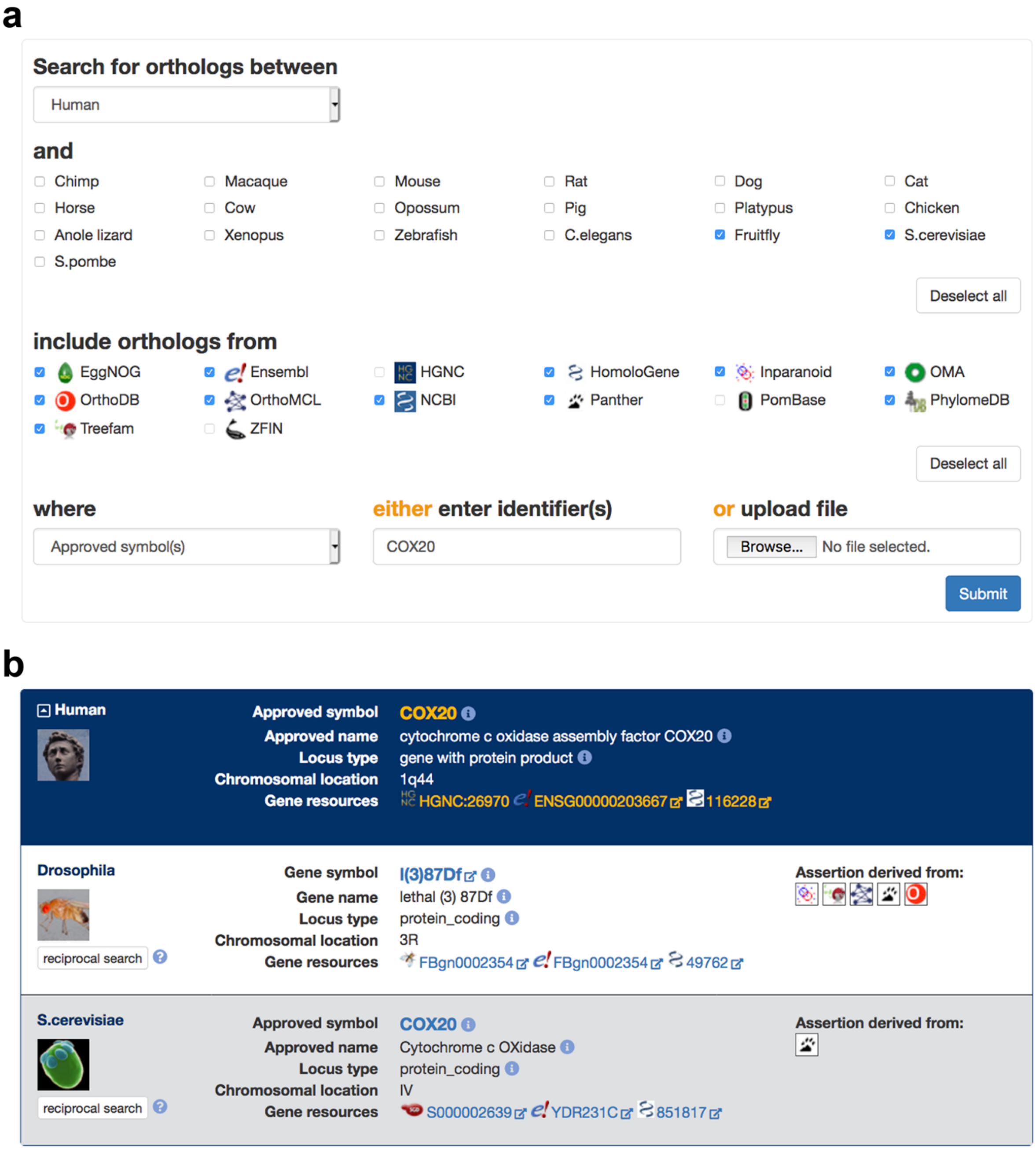
TreeFam gene tree for family awaiting annotation. You can download the tree here.

How to cite this tree +

Control elements +

Tree type: ● Model ○ Wormbase ○ Full
Branch length: ● Cladogram ○ Real
Leaf labels: ● Gene name ○ UniProt
Internal labels: □ Bootstrap □ Taxonomy
Mouse Zooming: ○ On ● Off

Figure 7



a

Search for orthologs between

Human

and

☐ Chimp ☐ Macaque ☐ Mouse ☐ Rat ☐ Dog ☐ Cat
☐ Horse ☐ Cow ☐ Opossum ☐ Pig ☐ Platypus ☐ Chicken
☐ Anole lizard ☐ Xenopus ☐ Zebrafish ☐ C.elegans ☑ Fruitfly ☑ S.cerevisiae
☐ S.pombe

Deselect all

include orthologs from

☑ EggNOG ☑ Ensembl ☐ HGNC ☑ HomoloGene ☑ Inparanoid ☑ OMA
☑ OrthoDB ☑ OrthoMCL ☑ NCBI ☑ Panther ☐ PomBase ☑ PhylomeDB
☑ Treefam ☐ ZFIN

Deselect all

where

Approved symbol(s)

either enter identifier(s)

COX20

or upload file

Browse... No file selected.

Submit

b

Human

Approved symbol: COX20
Approved name: cytochrome c oxidase assembly factor COX20
Locus type: gene with protein product
Chromosomal location: 1q44
Gene resources: HGNC:26970 ENSG00000203667 116228

Drosophila

reciprocal search

Gene symbol: l(3)87Df
Gene name: lethal (3) 87Df
Locus type: protein_coding
Chromosomal location: 3R
Gene resources: FBgn0002354 FBgn0002354 49762

Assertion derived from:

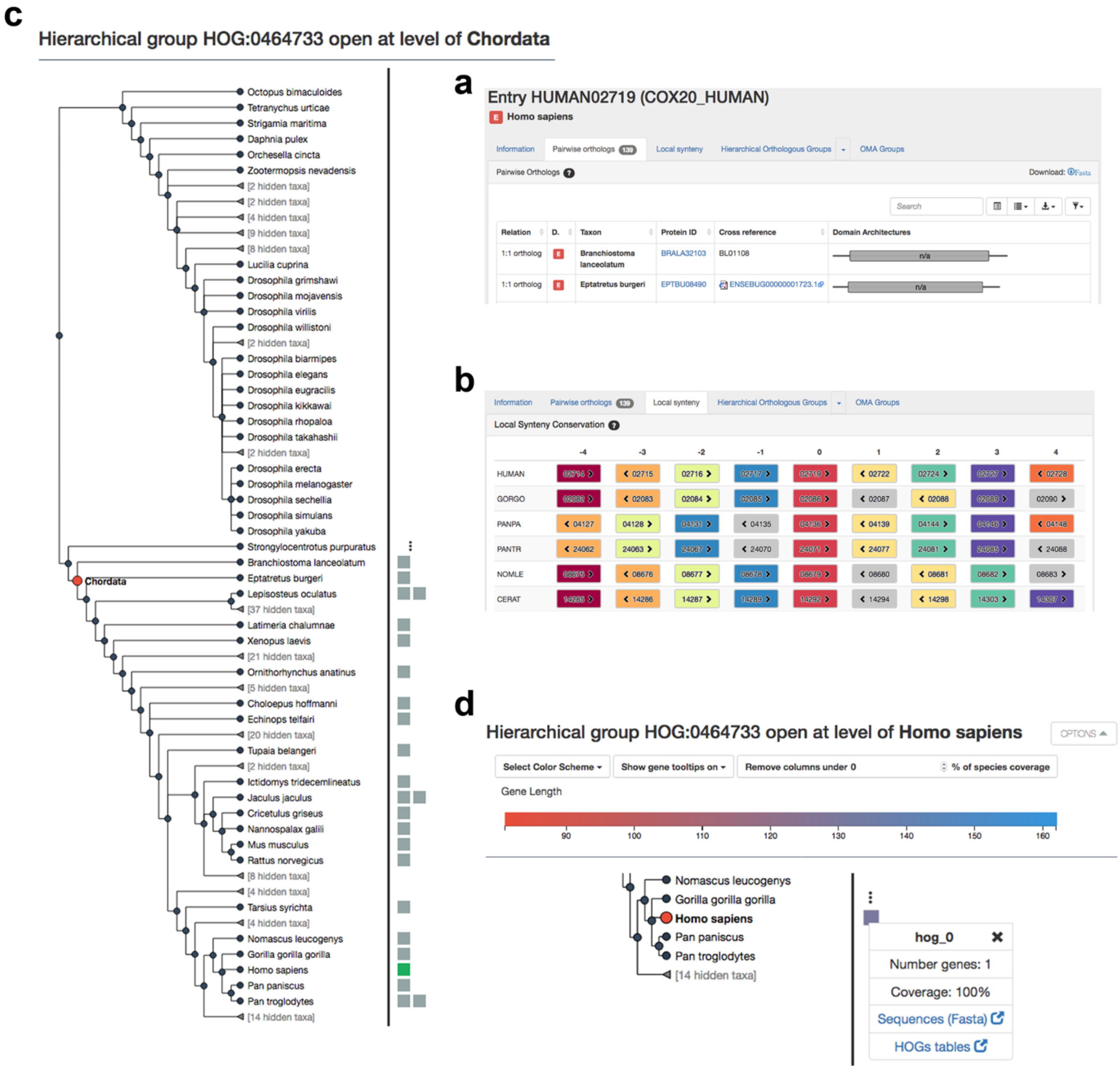
S.cerevisiae

reciprocal search

Approved symbol: COX20
Approved name: Cytochrome c OXidase
Locus type: protein_coding
Chromosomal location: IV
Gene resources: S000002639 YDR231C 851817

Assertion derived from:

Figure 8



Entry HUMAN02719 (COX20_HUMAN)

E Homo sapiens

Pairwise Orthologs

Relation	D.	Taxon	Protein ID	Cross reference	Domain Architectures
1:1 ortholog	E	Branchiostoma lanceolatum	BRALA32103	BL01108	n/a
1:1 ortholog	E	Eptatretus burgeri	EPTBU08490	ENSEBUG00000001723.1	n/a

Local Synteny Conservation

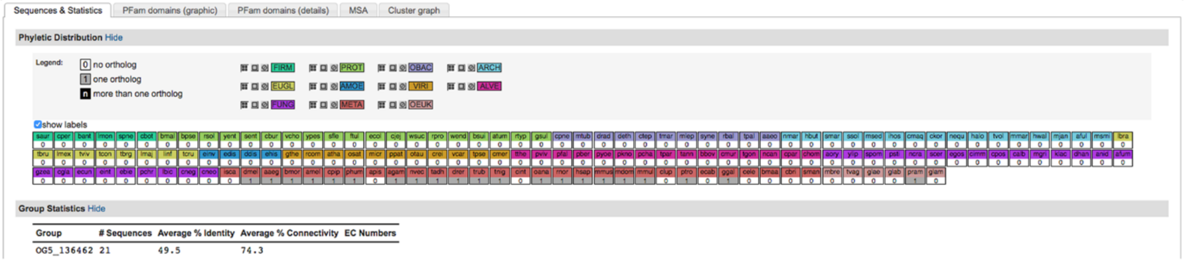
	-4	-3	-2	-1	0	1	2	3	4
HUMAN	02714 ›	‹ 02715	02716 ›	02717 ›	02719 ›	‹ 02722	02724 ›	02727 ›	‹ 02728
GORGO	02082 ›	‹ 02083	02084 ›	02085 ›	02086 ›	‹ 02087	‹ 02088	02089 ›	02090 ›
PANPA	‹ 04127	04128 ›	04131 ›	‹ 04135	04136 ›	‹ 04139	04144 ›	04146 ›	‹ 04148
PANTR	‹ 24062	24063 ›	24067 ›	‹ 24070	24071 ›	‹ 24077	24081 ›	24085 ›	‹ 24088
NOMLE	08675 ›	‹ 08676	08677 ›	08678 ›	08679 ›	‹ 08680	‹ 08681	08682 ›	08683 ›
CERAT	14285 ›	‹ 14286	14287 ›	14289 ›	14292 ›	‹ 14294	‹ 14298	14303 ›	14307 ›

Hierarchical group HOG:0464733 open at level of Homo sapiens

hog_0
Number genes: 1
Coverage: 100%
Sequences (Fasta)
HOGs tables

Figure 9

a



Phyletic Distribution Hide

Legend:
0 no ortholog
1 one ortholog
n more than one ortholog

FIRM PROT OBAC ARCH
EUGL AMOE VIRI ALVE
FUNG META OEUK

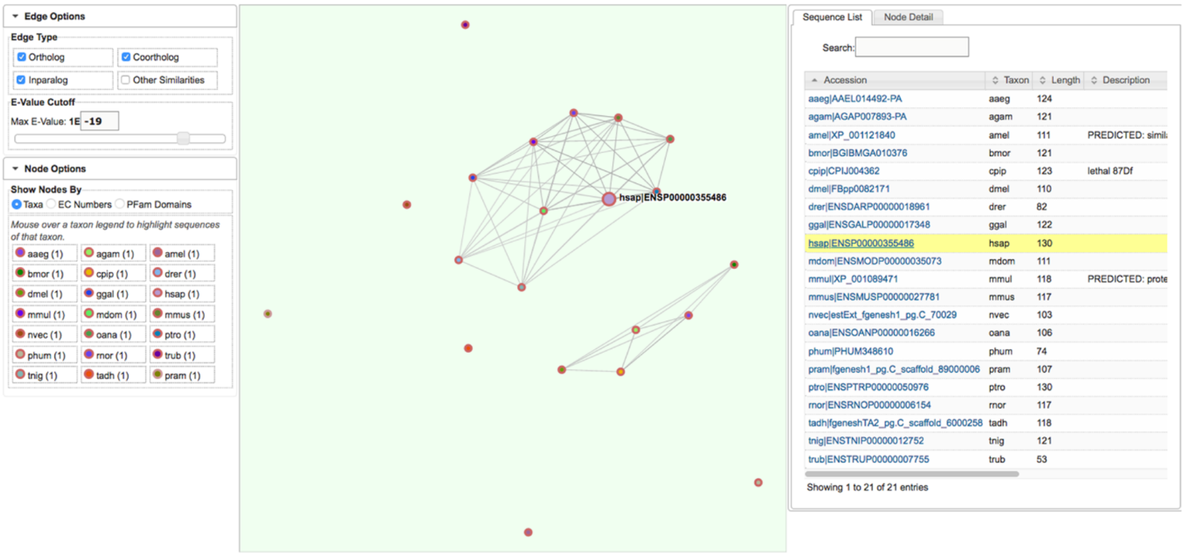
☑show labels

saur	cper	bant	lmon	spne	cbot	bmal	bpse	rsol	yent	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmar	mlep	syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed	ihos	cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	lbra
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tbru	lmex	tviv	tcon	tbrg	lmaj	linf	tcru	einv	edis	ddis	ehis	gthe	rcom	atha	osat	micr	ppat	otau	crei	vcar	tpse	cmer	tthe	pviv	pfal	pber	pyoe	pkno	pcha	tpar	tann	bbov	cmur	tgon	ncan	cpar	chom	aory	ylip	spom	psti	ncra	scer	egos	cimm	cpos	calb	mgri	klac	dhan	anid	afum
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gzea	cgla	ecun	eint	ebie	pchr	lbic	cneg	cneo	isca	dmel	aaeg	bmor	amel	cpip	phum	apis	agam	nvec	tadh	drer	trub	tnig	cint	oana	rnor	hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam									
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0									

Group Statistics Hide

Group	# Sequences	Average % Identity	Average % Connectivity	EC Numbers
OG5_136462	21	49.5	74.3	

b



▾ **Edge Options**

Edge Type

☑ Ortholog ☑ Coortholog
☑ Inparalog ☐ Other Similarities

E-Value Cutoff

Max E-Value: **1E -19**

▾ **Node Options**

Show Nodes By

● Taxa ○ EC Numbers ○ PFam Domains

Mouse over a taxon legend to highlight sequences of that taxon.

● aaeg (1)	● agam (1)	● amel (1)
● bmor (1)	● cpip (1)	● drer (1)
● dmel (1)	● ggal (1)	● hsap (1)
● mmul (1)	● mdom (1)	● mmus (1)
● nvec (1)	● oana (1)	● ptro (1)
● phum (1)	● rnor (1)	● trub (1)
● tnig (1)	● tadh (1)	● pram (1)

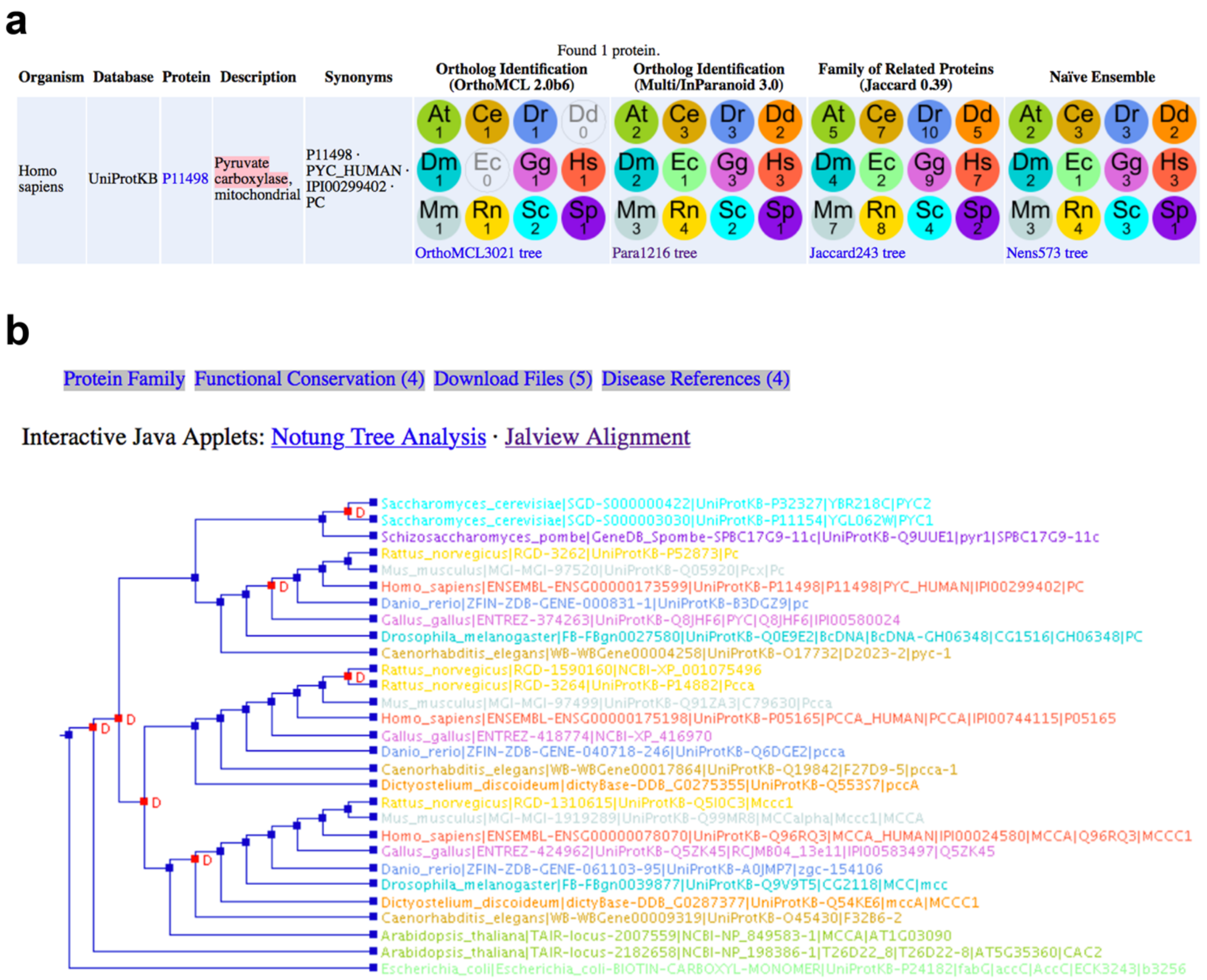
Sequence List | Node Detail

Search:

Accession	Taxon	Length	Description
aaeg|AAEL014492-PA	aaeg	124	
agam|AGAP007893-PA	agam	121	
amel|XP_001121840	amel	111	PREDICTED: simil
bmor|BGIBMGA010376	bmor	121	
cpip|CPIJ004362	cpip	123	lethal 87Df
dmel|FBpp0082171	dmel	110	
drer|ENSDARP00000018961	drer	82	
ggal|ENSGALP00000017348	ggal	122	
hsap|ENSP00000355486	hsap	130	
mdom|ENSMODP00000035073	mdom	111	
mmul|XP_001089471	mmul	118	PREDICTED: prote
mmus|ENSMUSP00000027781	mmus	117	
nvec|estExt_fgenesh1_pg.C_70029	nvec	103	
oana|ENSOANP00000016266	oana	106	
phum|PHUM348610	phum	74	
pram|fgenesh1_pg.C_scaffold_89000006	pram	107	
ptro|ENSPTRP00000050976	ptro	130	
rnor|ENSRNOP00000006154	rnor	117	
tadh|fgeneshTA2_pg.C_scaffold_6000258	tadh	118	
tnig|ENSTNIP00000012752	tnig	121	
trub|ENSTRUP00000007755	trub	53	

Showing 1 to 21 of 21 entries

Figure 10



Found 1 protein.

Organism	Database	Protein	Description	Synonyms	Ortholog Identification (OrthoMCL 2.0b6)	Ortholog Identification (Multi/InParanoid 3.0)	Family of Related Proteins (Jaccard 0.39)	Naïve Ensemble
Homo sapiens	UniProtKB	P11498	Pyruvate carboxylase, mitochondrial	P11498 · PYC_HUMAN · IPI00299402 · PC	At 1 · Ce 1 · Dr 1 · Dd 0 Dm 1 · Ec 0 · Gg 1 · Hs 1 Mm 1 · Rn 1 · Sc 2 · Sp 1 OrthoMCL3021 tree	At 2 · Ce 3 · Dr 3 · Dd 2 Dm 2 · Ec 1 · Gg 3 · Hs 3 Mm 3 · Rn 4 · Sc 2 · Sp 1 Para1216 tree	At 5 · Ce 7 · Dr 10 · Dd 5 Dm 4 · Ec 2 · Gg 9 · Hs 7 Mm 7 · Rn 8 · Sc 4 · Sp 2 Jaccard243 tree	At 2 · Ce 3 · Dr 3 · Dd 2 Dm 2 · Ec 1 · Gg 3 · Hs 3 Mm 3 · Rn 4 · Sc 3 · Sp 1 Nens573 tree

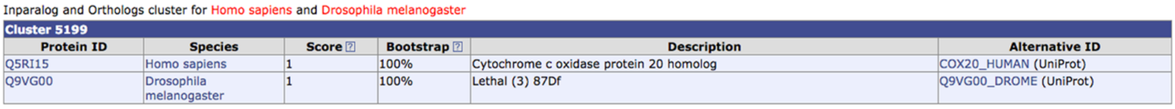
Protein Family Functional Conservation (4) Download Files (5) Disease References (4)

Interactive Java Applets: Notung Tree Analysis · Jalview Alignment

```
Saccharomyces_cerevisiae|SGD-S000000422|UniProtKB-P32327|YBR218C|PYC2
Saccharomyces_cerevisiae|SGD-S000003030|UniProtKB-P11154|YGL062W|PYC1
Schizosaccharomyces_pombe|GeneDB_Spombe-SPBC17G9-11c|UniProtKB-Q9UUE1|pyr1|SPBC17G9-11c
Rattus_norvegicus|RGD-3262|UniProtKB-P52873|Pc
Mus_musculus|MGI-MGI-97520|UniProtKB-Q05920|Pcx|Pc
Homo_sapiens|ENSEMBL-ENSG00000173599|UniProtKB-P11498|P11498|PYC_HUMAN|IPI00299402|PC
Danio_rerio|ZFIN-ZDB-GENE-000831-1|UniProtKB-B3DGZ9|pc
Gallus_gallus|ENTREZ-374263|UniProtKB-Q8JHF6|PYC|Q8JHF6|IPI00580024
Drosophila_melanogaster|FB-FBgn0027580|UniProtKB-Q0E9E2|BcDNA|BcDNA-GH06348|CG1516|GH06348|PC
Caenorhabditis_elegans|WB-WBGene00004258|UniProtKB-O17732|D2023-2|pyc-1
Rattus_norvegicus|RGD-1590160|NCBI-XP_001075496
Rattus_norvegicus|RGD-3264|UniProtKB-P14882|Pcca
Mus_musculus|MGI-MGI-97499|UniProtKB-Q91ZA3|C79630|Pcca
Homo_sapiens|ENSEMBL-ENSG00000175198|UniProtKB-P05165|PCCA_HUMAN|PCCA|IPI00744115|P05165
Gallus_gallus|ENTREZ-418774|NCBI-XP_416970
Danio_rerio|ZFIN-ZDB-GENE-040718-246|UniProtKB-Q6DGE2|pcca
Caenorhabditis_elegans|WB-WBGene00017864|UniProtKB-Q19842|F27D9-5|pcca-1
Dictyostelium_discoideum|dictyBase-DDB_G0275355|UniProtKB-Q553S7|pccA
Rattus_norvegicus|RGD-1310615|UniProtKB-Q5I0C3|Mccc1
Mus_musculus|MGI-MGI-1919289|UniProtKB-Q99MR8|MCCalpha|Mccc1|MCCA
Homo_sapiens|ENSEMBL-ENSG00000078070|UniProtKB-Q96RQ3|MCCA_HUMAN|IPI00024580|MCCA|Q96RQ3|MCCC1
Gallus_gallus|ENTREZ-424962|UniProtKB-Q5ZK45|RCJMB04_13e11|IPI00583497|Q5ZK45
Danio_rerio|ZFIN-ZDB-GENE-061103-95|UniProtKB-A0JMP7|zgc-154106
Drosophila_melanogaster|FB-FBgn0039877|UniProtKB-Q9V9T5|CG2118|MCC|mcc
Dictyostelium_discoideum|dictyBase-DDB_G0287377|UniProtKB-Q54KE6|mccA|MCCC1
Caenorhabditis_elegans|WB-WBGene00009319|UniProtKB-O45430|F32B6-2
Arabidopsis_thaliana|TAIR-locus-2007559|NCBI-NP_849583-1|MCCA|AT1G03090
Arabidopsis_thaliana|TAIR-locus-2182658|NCBI-NP_198386-1|T26D22_8|T26D22-8|AT5G35360|CAC2
Escherichia_coli|Escherichia_coli-BIOTIN-CARBOXYL-MONOMER|UniProtKB-P24182|fabG|accC|AccC|ECK3243|b3256
```

Figure 11

a

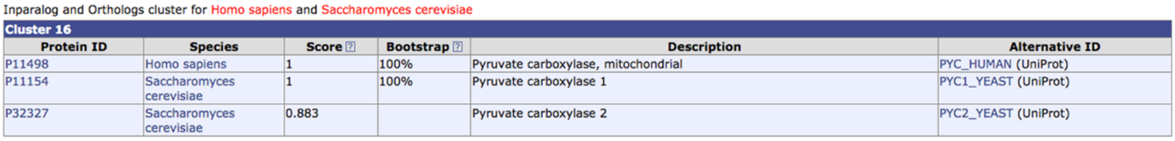


Inparalog and Orthologs cluster for Homo sapiens and Drosophila melanogaster

Cluster 5199

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5RI15	Homo sapiens	1	100%	Cytochrome c oxidase protein 20 homolog	COX20_HUMAN (UniProt)
Q9VG00	Drosophila melanogaster	1	100%	Lethal (3) 87Df	Q9VG00_DROME (UniProt)

b

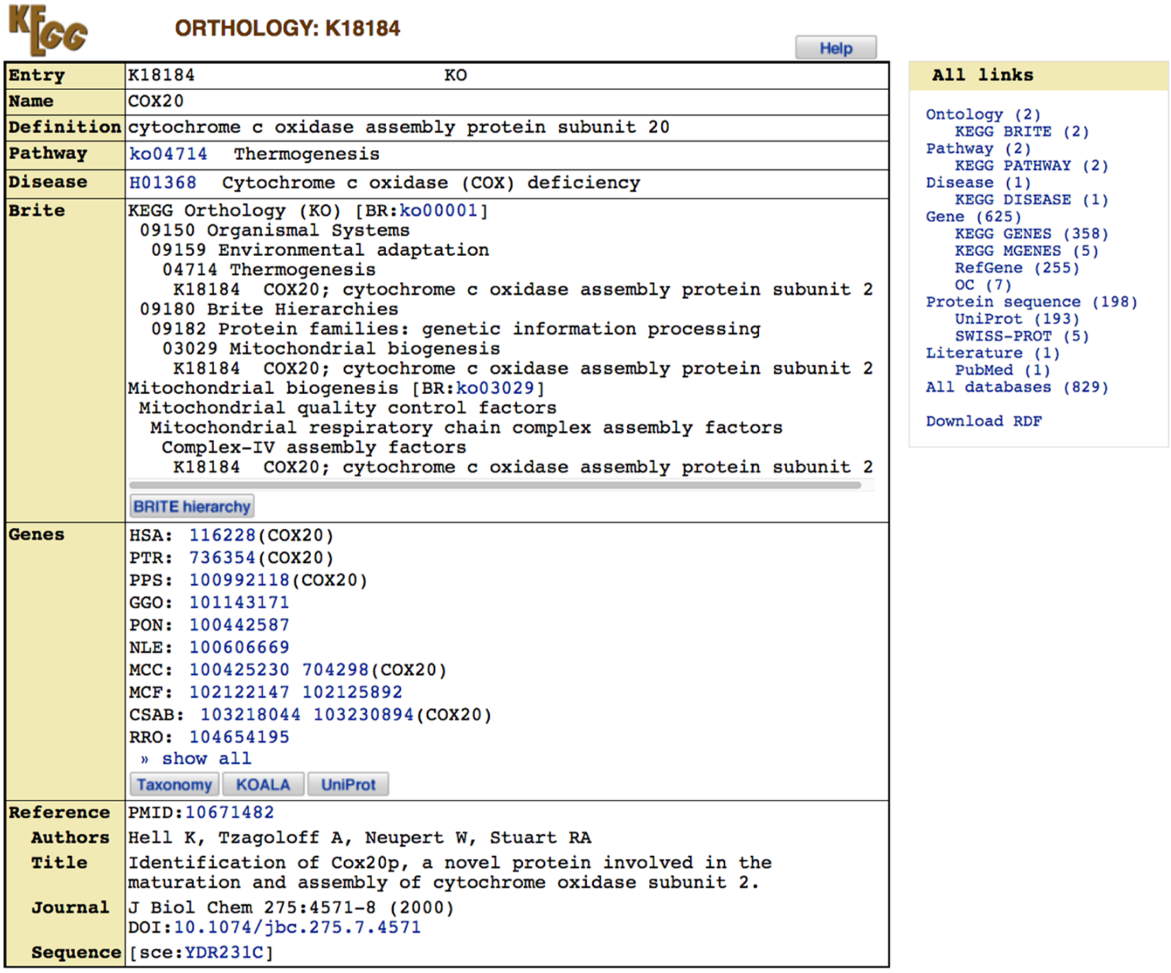


Inparalog and Orthologs cluster for Homo sapiens and Saccharomyces cerevisiae

Cluster 16

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
P11498	Homo sapiens	1	100%	Pyruvate carboxylase, mitochondrial	PYC_HUMAN (UniProt)
P11154	Saccharomyces cerevisiae	1	100%	Pyruvate carboxylase 1	PYC1_YEAST (UniProt)
P32327	Saccharomyces cerevisiae	0.883		Pyruvate carboxylase 2	PYC2_YEAST (UniProt)

c



ORTHOLOGY: K18184

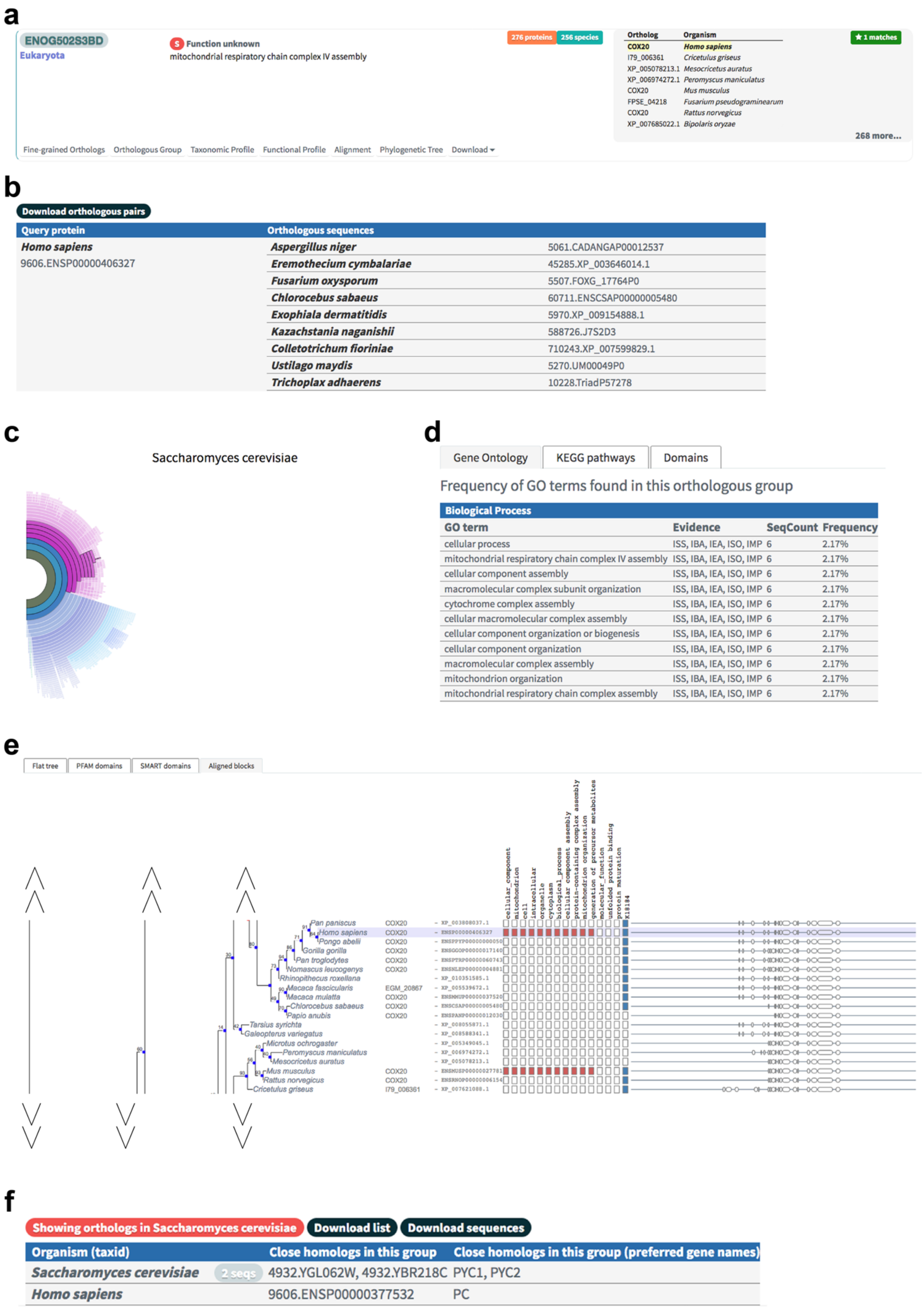
Entry	K18184 KO
Name	COX20
Definition	cytochrome c oxidase assembly protein subunit 20
Pathway	ko04714 Thermogenesis
Disease	H01368 Cytochrome c oxidase (COX) deficiency
Brite	KEGG Orthology (KO) [BR:ko00001] 09150 Organismal Systems 09159 Environmental adaptation 04714 Thermogenesis K18184 COX20; cytochrome c oxidase assembly protein subunit 2 09180 Brite Hierarchies 09182 Protein families: genetic information processing 03029 Mitochondrial biogenesis K18184 COX20; cytochrome c oxidase assembly protein subunit 2 Mitochondrial biogenesis [BR:ko03029] Mitochondrial quality control factors Mitochondrial respiratory chain complex assembly factors Complex-IV assembly factors K18184 COX20; cytochrome c oxidase assembly protein subunit 2 BRITE hierarchy
Genes	HSA: 116228(COX20) PTR: 736354(COX20) PPS: 100992118(COX20) GGO: 101143171 PON: 100442587 NLE: 100606669 MCC: 100425230 704298(COX20) MCF: 102122147 102125892 CSAB: 103218044 103230894(COX20) RRO: 104654195 » show all Taxonomy KOALA UniProt
Reference	PMID:10671482
Authors	Hell K, Tzagoloff A, Neupert W, Stuart RA
Title	Identification of Cox20p, a novel protein involved in the maturation and assembly of cytochrome oxidase subunit 2.
Journal	J Biol Chem 275:4571-8 (2000) DOI:10.1074/jbc.275.7.4571
Sequence	[sce:YDR231C]

All links

- Ontology (2)
  - KEGG BRITE (2)
- Pathway (2)
  - KEGG PATHWAY (2)
- Disease (1)
  - KEGG DISEASE (1)
- Gene (625)
  - KEGG GENES (358)
  - KEGG MGENES (5)
  - RefGene (255)
  - OC (7)
- Protein sequence (198)
  - UniProt (193)
  - SWISS-PROT (5)
- Literature (1)
  - PubMed (1)
- All databases (829)

Download RDF

Figure 12



Ortholog	Organism
COX20	*Homo sapiens*
I79_006361	*Cricetulus griseus*
XP_005078213.1	*Mesocricetus auratus*
XP_006974272.1	*Peromyscus maniculatus*
COX20	*Mus musculus*
FPSE_04218	*Fusarium pseudograminearum*
COX20	*Rattus norvegicus*
XP_007685022.1	*Bipolaris oryzae*

Query protein	Orthologous sequences	
Homo sapiens 9606.ENSP00000406327	***Aspergillus niger***	5061.CADANGAP00012537
	Eremothecium cymbalariae	45285.XP_003646014.1
	Fusarium oxysporum	5507.FOXG_17764P0
	Chlorocebus sabaeus	60711.ENSCSAP00000005480
	Exophiala dermatitidis	5970.XP_009154888.1
	Kazachstania naganishii	588726.J7S2D3
	Colletotrichum fioriniae	710243.XP_007599829.1
	Ustilago maydis	5270.UM00049P0
	Trichoplax adhaerens	10228.TriadP57278

Frequency of GO terms found in this orthologous group

Biological Process

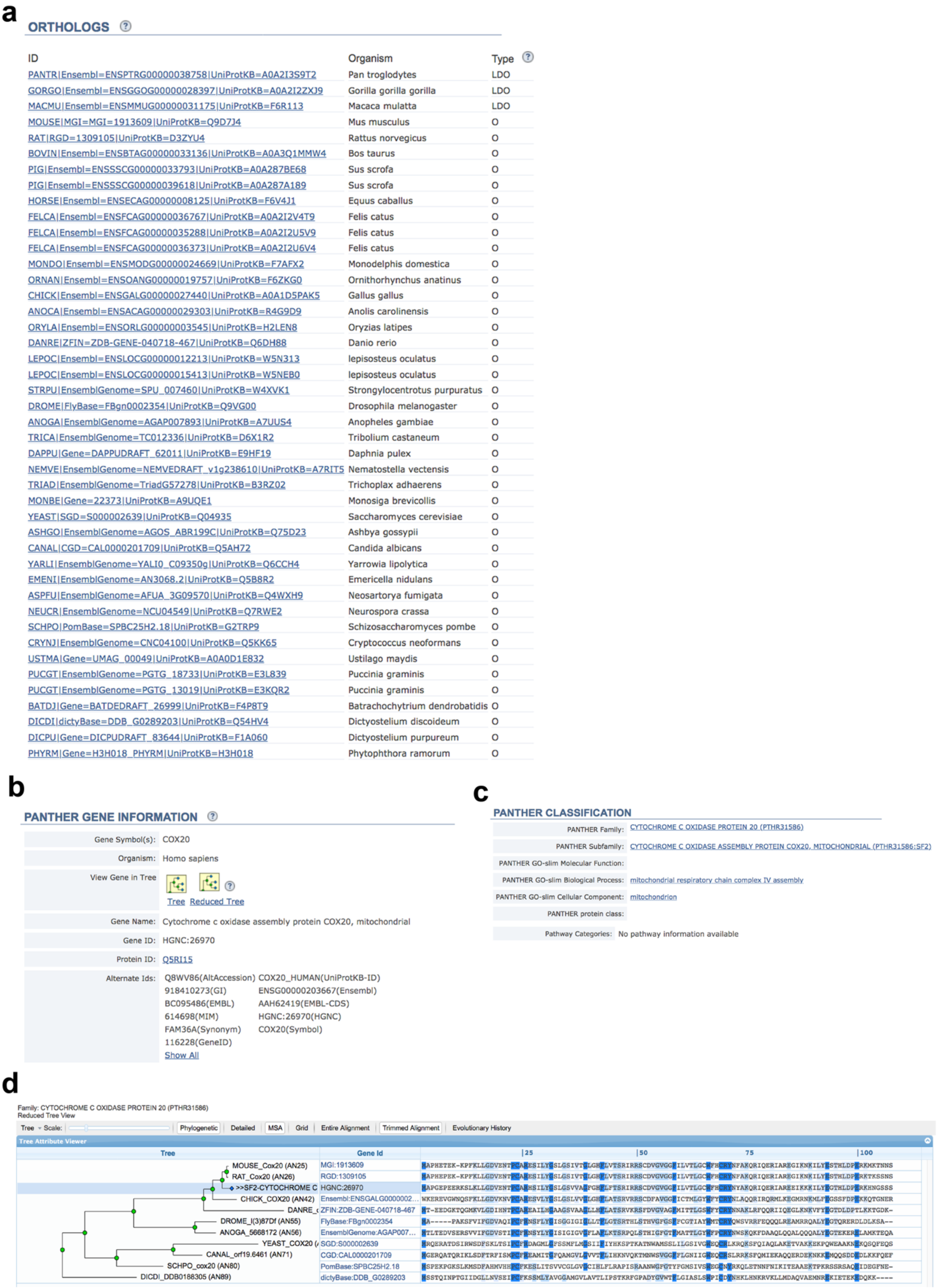
GO term	Evidence	SeqCount	Frequency
cellular process	ISS, IBA, IEA, ISO, IMP	6	2.17%
mitochondrial respiratory chain complex IV assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%
cellular component assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%
macromolecular complex subunit organization	ISS, IBA, IEA, ISO, IMP	6	2.17%
cytochrome complex assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%
cellular macromolecular complex assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%
cellular component organization or biogenesis	ISS, IBA, IEA, ISO, IMP	6	2.17%
cellular component organization	ISS, IBA, IEA, ISO, IMP	6	2.17%
macromolecular complex assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%
mitochondrion organization	ISS, IBA, IEA, ISO, IMP	6	2.17%
mitochondrial respiratory chain complex assembly	ISS, IBA, IEA, ISO, IMP	6	2.17%

e
Flat tree
PFAM domains
SMART domains
Aligned blocks
cellular_component
mitochondrion
cell
intracellular
organelle
cytoplasm
biological_process
cellular component assembly
protein-containing complex assembly
mitochondrion organization
generation of precursor metabolites
molecular_function
unfolded protein binding
protein maturation
K18184
Pan paniscus — XP_003808037.1
Homo sapiens COX20 — ENSP00000406327
Pongo abelii COX20 — ENSPPYP00000000050
Gorilla gorilla COX20 — ENSGGOP00000017140
Pan troglodytes COX20 — ENSPTRP00000060743
Nomascus leucogenys COX20 — ENSNLEP00000004881
Rhinopithecus roxellana — XP_010351585.1
Macaca fascicularis EGM_20867 — XP_005539672.1
Macaca mulatta COX20 — ENSMMUP00000037520
Chlorocebus sabaeus COX20 — ENSCSAP00000005480
Papio anubis COX20 — ENSPANP00000012030
Tarsius syrichta — XP_008055871.1
Galeopterus variegatus — XP_008588341.1
Microtus ochrogaster — XP_005349045.1
Peromyscus maniculatus — XP_006974272.1
Mesocricetus auratus — XP_005078213.1
Mus musculus COX20 — ENSMUSP00000027781
Rattus norvegicus COX20 — ENSRNOP00000006154
Cricetulus griseus I79_006361 — XP_007621088.1
91
84
71
86
94
73
90
49
70
80
30
42
14
40
40
56
93
93
f

Showing orthologs in Saccharomyces cerevisiae | Download list | Download sequences

Organism (taxid)	Close homologs in this group	Close homologs in this group (preferred gene names)
Saccharomyces cerevisiae 2 seqs	4932.YGL062W, 4932.YBR218C	PYC1, PYC2
Homo sapiens	9606.ENSP00000377532	PC

Figure 13



ORTHOLOGS

ID	Organism	Type
PANTR|Ensembl=ENSPTRG00000038758|UniProtKB=A0A2I3S9T2	Pan troglodytes	LDO
GORGO|Ensembl=ENSGGOG00000028397|UniProtKB=A0A2I2ZXJ9	Gorilla gorilla gorilla	LDO
MACMU|Ensembl=ENSMMUG00000031175|UniProtKB=F6R113	Macaca mulatta	LDO
MOUSE|MGI=MGI=1913609|UniProtKB=Q9D7J4	Mus musculus	O
RAT|RGD=1309105|UniProtKB=D3ZYU4	Rattus norvegicus	O
BOVIN|Ensembl=ENSBTAG00000033136|UniProtKB=A0A3Q1MMW4	Bos taurus	O
PIG|Ensembl=ENSSSCG00000033793|UniProtKB=A0A287BE68	Sus scrofa	O
PIG|Ensembl=ENSSSCG00000039618|UniProtKB=A0A287A189	Sus scrofa	O
HORSE|Ensembl=ENSECAG00000008125|UniProtKB=F6V4J1	Equus caballus	O
FELCA|Ensembl=ENSFCAG00000036767|UniProtKB=A0A2I2V4T9	Felis catus	O
FELCA|Ensembl=ENSFCAG00000035288|UniProtKB=A0A2I2U5V9	Felis catus	O
FELCA|Ensembl=ENSFCAG00000036373|UniProtKB=A0A2I2U6V4	Felis catus	O
MONDO|Ensembl=ENSMODG00000024669|UniProtKB=F7AFX2	Monodelphis domestica	O
ORNAN|Ensembl=ENSOANG00000019757|UniProtKB=F6ZKG0	Ornithorhynchus anatinus	O
CHICK|Ensembl=ENSGALG00000027440|UniProtKB=A0A1D5PAK5	Gallus gallus	O
ANOCA|Ensembl=ENSACAG00000029303|UniProtKB=R4G9D9	Anolis carolinensis	O
ORYLA|Ensembl=ENSORLG00000003545|UniProtKB=H2LEN8	Oryzias latipes	O
DANRE|ZFIN=ZDB-GENE-040718-467|UniProtKB=Q6DH88	Danio rerio	O
LEPOC|Ensembl=ENSLOCG00000012213|UniProtKB=W5N313	lepisosteus oculatus	O
LEPOC|Ensembl=ENSLOCG00000015413|UniProtKB=W5NEB0	lepisosteus oculatus	O
STRPU|EnsemblGenome=SPU_007460|UniProtKB=W4XVK1	Strongylocentrotus purpuratus	O
DROME|FlyBase=FBgn0002354|UniProtKB=Q9VG00	Drosophila melanogaster	O
ANOGA|EnsemblGenome=AGAP007893|UniProtKB=A7UUS4	Anopheles gambiae	O
TRICA|EnsemblGenome=TC012336|UniProtKB=D6X1R2	Tribolium castaneum	O
DAPPU|Gene=DAPPUDRAFT_62011|UniProtKB=E9HF19	Daphnia pulex	O
NEMVE|EnsemblGenome=NEMVEDRAFT_v1g238610|UniProtKB=A7RIT5	Nematostella vectensis	O
TRIAD|EnsemblGenome=TriadG57278|UniProtKB=B3RZ02	Trichoplax adhaerens	O
MONBE|Gene=22373|UniProtKB=A9UQE1	Monosiga brevicollis	O
YEAST|SGD=S000002639|UniProtKB=Q04935	Saccharomyces cerevisiae	O
ASHGO|EnsemblGenome=AGOS_ABR199C|UniProtKB=Q75D23	Ashbya gossypii	O
CANAL|CGD=CAL0000201709|UniProtKB=Q5AH72	Candida albicans	O
YARLI|EnsemblGenome=YALI0_C09350g|UniProtKB=Q6CCH4	Yarrowia lipolytica	O
EMENI|EnsemblGenome=AN3068.2|UniProtKB=Q5B8R2	Emericella nidulans	O
ASPFU|EnsemblGenome=AFUA_3G09570|UniProtKB=Q4WXH9	Neosartorya fumigata	O
NEUCR|EnsemblGenome=NCU04549|UniProtKB=Q7RWE2	Neurospora crassa	O
SCHPO|PomBase=SPBC25H2.18|UniProtKB=G2TRP9	Schizosaccharomyces pombe	O
CRYNJ|EnsemblGenome=CNC04100|UniProtKB=Q5KK65	Cryptococcus neoformans	O
USTMA|Gene=UMAG_00049|UniProtKB=A0A0D1E832	Ustilago maydis	O
PUCGT|EnsemblGenome=PGTG_18733|UniProtKB=E3L839	Puccinia graminis	O
PUCGT|EnsemblGenome=PGTG_13019|UniProtKB=E3KQR2	Puccinia graminis	O
BATDJ|Gene=BATDEDRAFT_26999|UniProtKB=F4P8T9	Batrachochytrium dendrobatidis	O
DICDI|dictyBase=DDB_G0289203|UniProtKB=Q54HV4	Dictyostelium discoideum	O
DICPU|Gene=DICPUDRAFT_83644|UniProtKB=F1A060	Dictyostelium purpureum	O
PHYRM|Gene=H3H018_PHYRM|UniProtKB=H3H018	Phytophthora ramorum	O

PANTHER GENE INFORMATION

Gene Symbol(s):	COX20
Organism:	Homo sapiens
View Gene in Tree	Tree Reduced Tree
Gene Name:	Cytochrome c oxidase assembly protein COX20, mitochondrial
Gene ID:	HGNC:26970
Protein ID:	Q5RI15

Alternate Ids:	Q8WV86(AltAccession) COX20_HUMAN(UniProtKB-ID)
	918410273(GI) ENSG00000203667(Ensembl)
	BC095486(EMBL) AAH62419(EMBL-CDS)
	614698(MIM) HGNC:26970(HGNC)
	FAM36A(Synonym) COX20(Symbol)
	116228(GeneID)
	Show All

PANTHER CLASSIFICATION

PANTHER Family:	CYTOCHROME C OXIDASE PROTEIN 20 (PTHR31586)
PANTHER Subfamily:	CYTOCHROME C OXIDASE ASSEMBLY PROTEIN COX20, MITOCHONDRIAL (PTHR31586:SF2)
PANTHER GO-slim Molecular Function:	
PANTHER GO-slim Biological Process:	mitochondrial respiratory chain complex IV assembly
PANTHER GO-slim Cellular Component:	mitochondrion
PANTHER protein class:	
Pathway Categories:	No pathway information available

Family: CYTOCHROME C OXIDASE PROTEIN 20 (PTHR31586)
Reduced Tree View

Tree	Gene Id	|25 |50 75 |100
MOUSE_Cox20 (AN25)	MGI:1913609	MAPHETEK-KPFKLLGDVENTPCARESILYGSLGSIVTGLGHFLVTSRIRRSCDVGVGGFILVTLGCWFHCRYNFAKQRIQERIAREGIKNKILYESTHLDPERKMKTNNS
RAT_Cox20 (AN26)	RGD:1309105	MAPHEPEK-KPFKLLGDVENTPCARESILYGSLGSIVTGLGHFLVTSRIRRSCDVGVGGFILVTLGCWFHCRYNFAKQRIQERIAREGIKNKILYESTHLDPERKTKSSNS
>>SF2-CYTOCHROME C	HGNC:26970	MAPGEPEERKSLKLLGDVENTPCARHSILYGSLGSVVAGFGHFLFTSRIRRSCDVGVGGFILVTLGCWFHCRYNYAKQRIQERIAREEIKKKILYEGTHLDPERKHNGSSS
CHICK_COX20 (AN42)	Ensembl:ENSGALG0000002…	WKEREVGWNQSFKLLGDVKNVPCARESVLYGSLGSLVVGLGHFLATSRVRRSCDFAVGGFICTMLGYWFYCRYNLAQQRIRQRMLKEGMRNKMLFEGSSFDPEKKQTGNER
DANRE_(	ZFIN:ZDB-GENE-040718-467	MT-EEDGKTQGMKVLGDIHNTPCAREAILHGAAGSVAAGLLHFLATSRVKRSFDVGVAGFMITTLGSWFYCRYNNAKLRFQQRIIQEGLKNKVFYEGTDLDPTLKKTGDK-
DROME_l(3)87Df (AN55)	FlyBase:FBgn0002354	MA-----PAKSFVIFGDVAQIPCFRNSFLYGISGGIGIGLLTFLGTSRTHLSTHVGFGSFFCGTIAYWMTCRYQWSVRRFEQQQLREAMRRQALYEGTQRERDLDLKSA--
ANOGA_5668172 (AN56)	EnsemblGenome:AGAP007…	MTLTEDVSERSVVIFGDVSTIPCFRNSYLYGISIGIGVGFLAFLKTSRPQLSTHIGFGTFMATTLGYWFPCRYNWSKQKFDAAQLQQALQQQALYEGTEKERELAMKTEQA
YEAST_COX20 (/	SGD:S000002639	MRQERATDSIRWSDFSKLTSIPCFRDAGMLGFSSMFLMGSIIFIYHKSPTKATNWAMSSLILGSIVGWEQCRLKRQKSFQIAQLAKETVAKKEKPQWEAAKNEKQSQFEQS
CANAL_orf19.6461 (AN71)	CGD:CAL0000201709	MGERQATQRIKLSDFTRFISMPCFREAMITGFQAMGVLGVVTFLIHKNVQKTMNWSVGGFFLGNIIGWEQCRSLRRKSFQMIEKAQQERDEKNKKEMQQSDDEDLKKFQEF
SCHPO_cox20 (AN80)	PomBase:SPBC25H2.18	MSPEKPGKSLKMSDFAHMVHHPCFKESLITSVVCGLGVGCIHLFLRAPISRAANWGFGTYFGMSIVSWEGCNYRKQLETNNFNIKITEAAEKTPKRSSRSAQEIDEGFNE-
DICDI_DDB0188305 (AN89)	dictyBase:DDB_G0289203	MSSTQINPTGIIDGLLNVSEIPCFKSSMLYAVGGAMGVLAVTLIPSTKRFGPADYGVWTFLGIASLSWPICDYNHKLHNKRVKLLMDAQVAEMNREKIETEKEDQKK----